ABSTRACT

BIOLOGICAL SCIENCE

LOCKHART, EDWARD R. B.S. CLARK ATLANTA UNIVERSITY, 2005

NEURAL NETWORK ANALYSIS OF mRNA SECONDARY–STRUCTURE ACROSS TRANSCRIPTOMES

Former Advisor: Associate Professor Williams Seffens, Ph.D.

Current Advisor: Associate Professor Jaideep Chaudhary, Ph.D.

Dissertation dated December 2010

This study examines mRNAs of less than 5000 base pairs in size, to determine the effects of base composition on folding free energy. Statistical analysis between the native mRNA and its randomized sequences was conducted, and when comparing mRNAs in human, chimp, chicken, mouse, and several other transcriptomes, we found that the native mRNAs were more stable (greater negative free energy of folding). It has been found that when length and base composition are conserved, native mRNA sequences are more stable than random mRNA sequences. More stable folding conformations have greater negative free energy values. This negative bias in free energies can be statistically measured as a Z-score which normalizes for sequence length. In an effort to determine if sequence patterns correlate with secondary structure, a neural network (JavaNNS) was trained using three training sets (Negative-Z, Near Zero-Z, Positive-Z) separately to compare the effect of neural network learning from the folding characteristics of the gene sequences. The training sets were typically allowed to run for up to 100,000 generations, and the resulting sum square errors were periodically saved. We found that

the negative Z-score training set gives lower neural network sum square errors than the positive Z-score training set, and the Z-scores near zero have the highest training error. This indicates that there are more detectable sequence patterns in genes with more secondary structure than in genes exhibiting more positive Z-scores.

NEURAL NETWORK ANALYSIS OF mRNA SECONDARY–STRUCTURE ACROSS TRANSCRIPTOMES

A DISSERTATION IN BIOLOGY

SUBMITTED TO THE FACULTY OF CLARK ATLANTA UNIVERSITY

IN PARTIAL FULFULLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

BY

EDWARD RONALD LOCKHART, JR

DEPARTMENT OF BIOLOGICAL SCIENCES

ATLANTA, GEORGIA

DECEMBER 2010

ACKNOWLEDGEMENTS

First and for most I would like to give honor and grace to my Lord and Savior Jesus Christ. Nothing would be possible without your favor and mercy. Secondly, Edward R. Lockhart Sr. and Lashon Allen Lockhart, thank you for loving each other and ultimately creating me. I am truly grateful for your support. Elvera Lockhart thank you for being the bundle of energy that you are. Michelle Lockhart thank you for understanding and remaining patient as you always have. I love you unconditionally. Dr. William Seffens thanks for the insights and knowledge on many areas relating to science. And last but certainly not least Dr. Jaideep Chaudhary thank you so much for extending yourself even though your "plate was already full". I am privileged to have had the opportunity to engage in scientific discussion with you.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OR FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Deoxyribonucleic acids, though consisting of polymeric chains capable of taking up an infinite variety of individual configurations, have stable, double-stranded structure in aqueous solution (Doty, Boedtker et al. 1959). Likewise, synthetic homopolyribonucleotides spontaneously join together in certain pairs and triplets to form helical configurations (Doty, Boedtker et al. 1959). In all of these cases, the formation of unique and essentially complete conformations is a result of the most efficient way of allowing the maximum number of hydrogen bonds to form. Thus, chemical bonds are held in a unique secondary structure, or configuration, due to hydrogen bonds assisted to some extent by van der Waal's forces (Doty, Boedtker et al. 1959). With this data it was implied that all RNA molecules have the ability to form secondary structures and that a RNA molecule that lack secondary structure was "significant" (Gralla and DeLisi 1974). Arguments have been advanced that these structures are the result of evolutionary pressures or of requirements for unique functions such as virus packaging(Gralla and DeLisi 1974).

Most RNA molecules, including messenger RNA and transfer RNA, act as cellular intermediaries; that is, they convert the genetic information stored in DNA into the proteins that provide cells with structure and enable them to carry out metabolism (von Heijne, Nilsson et al. 1978). The primary structure of an RNA or DNA is simply the

5' to 3' list of covalently linked nucleotides, named by the attached base-base modifications usually described by the use of single-letter codes in place of the unmodified base (Zuker 2000). Helices and base pair stacking are inferred when two or more base pairs occur adjacent to one another. In fact, an important structural feature of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. The presence of this functional group causes the helix to adopt the A-form geometry rather than the B-form most commonly observed in DNA (Hermann and Patel 2000).

The wealth of nucleic acid sequences has prompted a search for secondary structures that may be of functional significance including transcription, replication, processing and regulation. Unfortunately, it is often difficult to know what the expected frequency of a particular secondary structure is due to chance alone (Fitch 1983). RNAs adopt specific secondary structures to carry out their biological functions, and exchange among alternative secondary structures plays essential roles in virtually all RNA-mediated processes ranging from RNA silencing and metabolite-activation of bacterial riboswitches to pre-mRNA splicing and viral RNA replication (Mahen, Watson et al. 2010). For example, transfer RNA (tRNA) adopts a defined secondary and tertiary structure (Holbrook 2005), whereas most messenger RNAs (mRNAs) exhibit only local structures (Graveley, Hertel et al. 1998). Defined RNA structures can be recognized by other molecules, as exemplified by aminoacyl-tRNA synthetases that contact with the minor groove of the tRNA acceptor stem (Sankaranarayanan, Dock-Bregeon et al. 1999), which illustrates the functional importance of RNA structure. The prevailing building blocks of large RNA molecules are double helices of base complementary strands.

Because of the limited number of packing arrangements, helices alone cannot account for the intriguing structural diversity observed in RNA structure (Hermann and Patel 2000). Because of these interactions, bulges stand out for their universal distribution in all types of structured functional RNAs. Bulges are unpaired stretches of nucleotides located within one strand of a nucleic acid duplex which is formed by hydrogen-bonded bases including canonical Watson–Crick and noncanonical base pairs. Bulge sizes can vary from a single unpaired residue up to several nucleotides that form frequently flexible extrusions from pseudo-continuous double helices (Hermann and Patel 2000). RNA structure is particularly suited for prediction because it is generally divided into two levels of complexity. The first level is secondary structure, involving canonical base-pairs. It is discrete in that each nucleotide is either paired or unpaired. The interactions that govern secondary structure are generally stronger than the interactions that determine the next level of structural complexity, tertiary structure, i.e. the three-dimensional shape (Banerjee et al., 1993; Jaeger et al., 1993; Laing & Draper, 1994; Crothers et al., 1974; Hilbers et al., 1976; Mathews et al., 1997).

Several studies have demonstrated that mRNA stability and secondary structure is an important factor in some gene expression systems (Rosenbaum, et.al., 1993; Kushner, 2002). Numerous mRNA-protein interactions are known to regulate gene expression including pre-mRNA splicing, polyadenylation, editing, transport, cytoplasmic targeting, translations and mRNA turnover (Sandberg and Mulroney, 2001). Many computational analyses of mRNA secondary structure have focused on 5' and 3' untranslated regions (UTRs), excluding the coding region (Shang et.al. 2004; Andrew et.al. 2004). Examples of important biological function involve particular secondary structures such as stem-

loops. Structural RNA features caused by complementary base-pairing are suspected to be involved in the regulation of mRNA degradation (Jacobson *et. al.* 1998). The classic example is the *trp-operon* of *E. coli* (Ramesh, 1993). These local structures are a small part of the overall global free energy of folding or minimum energy required to keep a molecule in its most stable conformation for the entire mRNA sequence. The majority of global free energy of folding arises from the coding sequence (CDS) and selection of codons (Seffens and Digby, 1999). Numerous statistical studies have established that codon frequencies are not random (Karlin and Brendel, 1993).

Hypothesis:

We hypothesize that the folding free energy (FFE) of mRNAs within transcriptomes is the result of specific sequence patterns in the genes. We will investigate this hypothesis by comparing various transcriptomes and their corresponding mRNA folding free energy Z-scores by using the pattern recognition capability of neural networks.

Specific Aims

Specific Aim 1: Compile Folding Free Energies of Transcriptomes.

- Conduct comparative study between the transcriptomes compiled using Z-score measures. We will calculate and analyze the folding free energies of native and randomized sequences.

- Examine complexity measures between transcriptomes to determine if secondary structure is related to animal evolutionary taxonomy. This analysis will be performed to assess if there is a phylogenetic trend that exist at the RNA level.

Specific Aim 2: Regression Analysis on mRNA sequences to determine what specific variables can be attributed to the differences in the folding free energies in species.

- Statistical analysis between native and randomized genes will be completed to display the statistical difference between the wild-type group (native) and the tested group (randomized)

- Base content and composition determination among the compiled species; specifically dinucleotide and trinucleotide content. We want to determine if base composition is correlated with secondary structure in genes characterized by Z-scores.

Specific Aim 3: Neural Network Analysis on mRNA sequences to detect specific sequence patterns.

- Determine if mRNA folding is related to the magnitude of Neural Network performance in detecting sequence patterns.

- Compare training sum squared errors to assess the complexity of patterns in codon usage in training sets.

These specific aims will us to determine if whether species contain native sequences that are classified as more stable than their randomized sequences. If so, then we believe transcriptomes are evolved systems that contain mRNAs in more stable conformations. We will also gain a better understanding as to how base composition and gene length relates to secondary structure categorized by Z-scores. Are species with larger gene lengths characterized as species with more secondary structure? Finally, completing these

specific aims will add comprehension to folding free energies and overall sequence patterns that may or may not be causing certain species to generate more or less secondary structure in their transcriptome. Analyzing the secondary structure present in transcriptomes and verifying pattern recognition associated with species could give more insight as why certain species have more secondary structure.

CHAPTER 2

REVIEW OF LITERATURE

Central Dogma

In the process of generating an organism several regulated processes must take place in order to ensure proper gene expression. Transcription and translation are two highly regulated steps involved in the central dogma (flow of information) which is the central foundation of molecular biology (figure 1). Both RNA and DNA use base pairs of nucleotides as complementary language that can be converted back and forth from DNA to RNA in the presence of the correct enzymes (Berg et al. 2006).

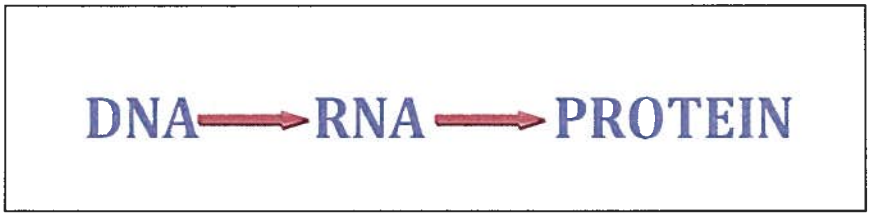

Figure 1.Central Dogma of Molecular Biology.

Structure of DNA

Deoxyribonucleic Acid or DNA contains an organism's hereditary information, including genetic instructions for the sequence of amino acids in polypeptiedes. DNA also carries the information needed to synthesize other nucleic acids such as ribonucleic acid, RNA. In living organisms, DNA does not usually exist as a single molecule, but

instead as a pair of molecules that are held tightly together (Watson and Crick 1953). The structure proposed by Watson and Crick consists of two polynucleotide chains wound helically around a common axis, tied together by hydrogen bonds between the purine and pyrimidine side chains (Delbruck 1954). The side chains of the two chains are arranged so that the purine adenine (A) is always matched with pyrimidine thymine (T) and the purine guanine (G) with pyrimidine cytosine (C). Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. The order, or sequence, of these bases determines the information available for building and maintaining an organism. Embodied in the structure of DNA as proposed by Watson and Crick was the suggestion that faithful duplication of base sequence was mediated primarily by hydrogen bonding between parental DNA and daughter nucleotides (Loeb, Springgate et al. 1974). The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. Each type of base on one strand forms a bond with just one type of base on the other strand, which is called complementary base pairing. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature (Gaub et al. 2000). The nucleotide repeats contain both the segment of the backbone and of the molecule, which holds the chain together, and a base, which interacts with the other DNA strand in the helix.

Transcription/RNA synthesis

Transcription, or RNA synthesis, is the process of creating an equivalent RNA copy from a sequence of DNA. RNA synthesis is catalyzed by a large enzyme called RNA polymerase. The basic biochemistry of RNA synthesis is common to prokaryotes and eukaryotes, although its regulation is more complex in eukaryotes. Despite substantial differences in size and number of polypeptide subunits, the overall structures of these enzymes are quite similar, revealing a common evolutionary origin (Berg et al 2002).

RNA synthesis, like nearly all biological polymerization reactions, takes place in three stages: initiation, elongation, and termination. RNA polymerase performs multiple functions in this process: the first step is initiation where RNA polymerase searches DNA for initiation sites, also called promoter regions. For instance, *E. coli* DNA has about 2000 promoter sites in its 4.8×10^6 bp genome (Berg et al 2002). Because these sequences are on the same molecule of DNA as the genes being transcribed, they are called cis-acting elements. Next, it unwinds a short stretch of double-helical DNA to produce a single-stranded DNA template from which it takes instructions (Berg et al 2002). Thirdly, RNA polymerase selects the correct ribonucleoside triphosphate and catalyzes the formation of a phosphodiester bond. The 3′-hydroxyl group of the last nucleotide in the chain nucleophilically attacks the α-phosphate group of the incoming nucleoside triphosphate with the concomitant release of a pyrophosphate (Berg et al 2002). This reaction is thermodynamically favorable, and the subsequent degradation of the pyrophosphate to orthophosphate locks the reaction in the direction of RNA synthesis (Berg et al 2002). This process is repeated many times as the enzyme moves

unidirectionally along the DNA template. Next, it detects termination signals that specify where a transcript ends. And finally RNA polymerase interacts with activator and repressor proteins that modulate the rate of transcription initiation over a wide dynamic range. These proteins, which play a more prominent role in eukaryotes than in prokaryotes, are called transcription factors or trans-acting elements.

Types of RNA

There are three major types of RNA, messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). The chemistry of RNA synthesis is identical for all forms of RNA (Berg et al 2002). Their synthetic process differs mainly in regulation, posttranscriptional processing, and the specific polymerase that participates (Berg et al 2002).

m-RNA

Messenger RNA is a single-stranded RNA molecule that is complementary to the template strand of DNA. mRNA is synthesized from a gene segment of DNA which ultimately contains the information on the primary sequence of amino acids in a polypeptide chain to be synthesized (Sussman, Holbrook et al. 1978). The messenger RNA is responsible for carrying the 'genetic code' into the cytoplasm where ribosomes are located to allow synthesis of protein.

t-RNA

Transfer RNA's are small single-stranded RNA's that contains on average about eighty (80) nucleotides, in which, three bases linked to form regions known as anticodons.

tRNA is responsible for "reading" the mRNA codon by using its own anticodon (figure 2). The actual "reading" is done by matching the base pairs through hydrogen bonding following the base pairing principle of Watson and Crick (Sussman, Holbrook et al. 1978). Each codon is "read" by various tRNA's until the appropriate match of the anticodon with the codon occurs. At the activation site, an activating enzyme (aminoacyl tRNA synthetase) adds a specific amino acid. Aminoacyl tRNA synthetases are important for two reasons first the attachment of a given amino acid to a particular tRNA establishes the genetic code. When an amino acid has been linked to a tRNA, it will be incorporated into a growing polypeptide chain at a position dictated by the anticodon of the tRNA (Berg et al 2002). Second, the formation of a peptide bond between free amino acids is not thermodynamically favorable (Berg et al 2002). The amino acid must first be activated or 'charged' for protein synthesis to proceed. The activated intermediates in protein synthesis are amino acid esters, in which the carboxyl group of an amino acid is linked to either the 2′- or the 3′-hydroxyl group of the ribose unit at the 3′ end of tRNA.

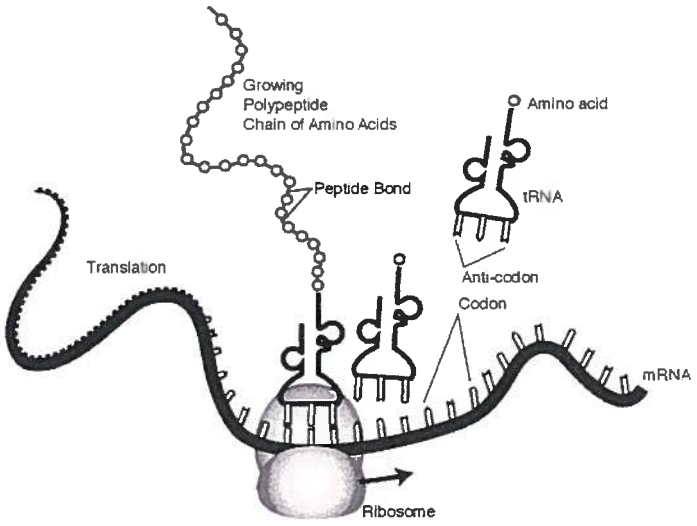

Figure 2. Translation Mechanism. Depiction of anticodons base pairing with codons to add amino acids to ultimately fold a protein.

r-RNA

The function of ribosomal RNA is to assist in decoding mRNA into amino acids and to interact with the tRNAs during translation by providing peptidyl transferase activity (allows enzyme to form peptide links to assist in transferring amino acids during translation) (Sussman, Holbrook et al. 1978) (figure 2). The ribosome is composed of two subunits, a large and small subunit. mRNA is sandwiched between the small and large subunits and the ribosome catalyzes the formation of a peptide bond between the 2 amino acids that are contained in the rRNA (figure 2). Prokaryotes contain a 70S ribosome with a large (50S) and small subunit (30S) (Pace 1973). The deproteinization of purified, bacterial ribosomes generally yields three distinct RNA components. Two of these, classified by their approximate sedimentation velocities as 23S and 5S rRNA's, are derived from the 50S ribosomal subunit (Rosset and Monier 1963). The third RNA component, 16S rRNA, originates from the 30S ribosomal subunit (Kurland 1960). Eukaryotes are comprised of 80S ribosomes which are composed of two subunits, one about 60S and a second about 40S in size (Taylor and Storck 1964). Deproteinization of the smaller, 40S subunit yields one RNA molecule, about 18S in size and 0.7 x 106 in molecular weight (Taylor, Glasgow et al. 1967). The 60S subunit yields three distinct RNA molecules. The largest of these is about 28S (Taylor, Glasgow et al. 1967).

RNA Structure

RNA is usually thought of as a single stranded linear molecule, however, in a biological system this is not always the case. Frequently, different regions of the same RNA strand will fold together via base pair interactions to form intricate secondary and

tertiary structures that are essential for correct biological function (Sussman 1978). Common secondary structure motifs include hairpin loops, stems, and bulges. Though RNA is usually single stranded, in some RNA virus genomes it will form a double stranded helix (Nowakowski, Shim et al. 1999). However, unlike DNA which form the B-form double helix, RNA forms an A-form double helix. The RNA double helix differs from that of the DNA double helix because of the presence of ribose, rather than deoxyribose, in the sugar phosphate backbone of the molecule. The addition of a hydroxyl group at the C2 postion in the ribose sugar is responsible for the A-form geometry in double stranded RNA (Szymanski 2000). The A-form makes a right-handed helix, like the B-form double helix of DNA, but is a shorter, wider helix than the B-form, and the major groove is deep, but narrow, making it virtually inaccessible to proteins (Szymanski 2000). It is in the major groove that the chemical groups are sequence-specific and dependent on base identity, and therefore, this is where proteins tend to bind a DNA double helix. Because the RNA A-form double helix contains a major groove that is too narrow and deep for proteins to access, the minor groove becomes more important for protein interactions with RNA helices (Szymanski 2000). Also, proteins that interact with specific RNA sequences commonly bind single-stranded RNA segments. When a strand of DNA forms a double helix with a strand of RNA, this will also result in an A-form helix (Szymanski 2000).

RNA Secondary Structure

The primary structure of a single-stranded nucleic acid, such as a tRNA, is the sequence of nucleotides or bases making up the molecule (Waterman 1978). Secondary structure of such a molecule is the interaction that exists between bases which preserves

the bonds in the primary structure but allows helical regions. When primary structure of a single-stranded tRNA is known, the question arises of which bases form pairs and allow the sequence to form helical regions in two dimensions (Waterman 1978). Secondary structure is defined as the structure of a nucleic acid or a protein that is created by the formation of hydrogen bonding between nucleotides and/or amino acids. Secondary structure has received much attention (Tinoco et al 1974) and has shown to have a role in the interactions of tRNAs with proteins (Pipas and McMahon 1975), in stabilizing mRNA, in packing RNA into virus particles, and in recognition of specific sites by components of the translating system (Lesk 1974).

Secondary Structure and mRNA Splicing

Several protein motifs that bind single stranded RNA have been characterized, and these are commonly found in splicing factors. Consistent with their action in a single-stranded state, a set of splicing enhancers and silencers has been confirmed bioinformatically to be more single stranded than bulk sequence, and to function more effectively when placed in the loop than the stem of a hairpin structure (Hiller, Zhang et al., 2007). The formation and stabilization of secondary structure around such regulatory elements is therefore a potential mechanism to reduce their effects on splicing (Smith, Query et al. 2008). More evidence that secondary structure is involved in gene regulation was conducted by investigators who reported that the secondary structure of a pre-mRNA influences a number of processing steps including alternative splicing (Hiller, Zhang et al. 2007). Since most splicing regulatory proteins bind to single-stranded RNA, the sequestration of RNA into double strands could prevent their binding. Hiller et al. analyzed the secondary structure context of experimentally determined splicing enhancer

and silencer motifs in their natural pre-mRNA context (Hiller, Zhang et al. 2007). They found that these splicing motifs are significantly more single-stranded than controls. These findings were validated by transfection experiments, where the effect of enhancer or silencer motifs on exon skipping was much more pronounced in single-stranded conformation (Hiller, Zhang et al., 2007). They also found that the structural context of predicted splicing motifs is under selection, suggesting a general importance of secondary structures on splicing and adding another level of evolutionary constraints on pre-mRNAs. These findings explain the action of mutations that affect splicing and indicate that the structural context of splicing motifs is part of the mRNA splicing code. Evidence suggests that splice sites themselves must be single stranded in order to allow spliceosome assembly, with secondary structure inhibiting U1 and U2 small nuclear ribonucleoproteins (snRNP) from binding (Liu, Goodall et al. 1995). Inclusion of the 3'splice sites in a hairpin is inhibitory for splicing, however this can be overcome by the presence of a single stranded ''helper'' downstream of 3'splice sites, likely recognized during assembly (Liu, Goodall et al. 1995). For example, the *S. cerevisiae* RP51B intron contains complementary sequences close to the 5' splice site and branch site that bring the ends of the intron together and aid spliceosome assembly (Charpentier and Rosbash, 1996), and it is possible that this is a common way to increase the efficiency of U1-U2 binding and intron definition.

Secondary Structure and Protein Synthesis

Structural elements of mRNA are known to play integral roles in mechanisms regulating translation and mRNA stability, which in turn directly affect translation efficiency and turnover rates of message, and therefore the amount of a specific protein

that is synthesized. For example, Wang and Wessler reported *Lc*, a member of the maize (*Zea mays*) *R/B* gene family, which encodes a basic helix-loop-helix transcriptional activator of the anthocyanin biosynthetic pathway, that a potential hairpin structure near the 5'end of the *Lc* mRNA (figure 3) also represses downstream translation in the rabbit reticulocyte in vitro translation system (Wang and Wessler 2001). Base pairing of the

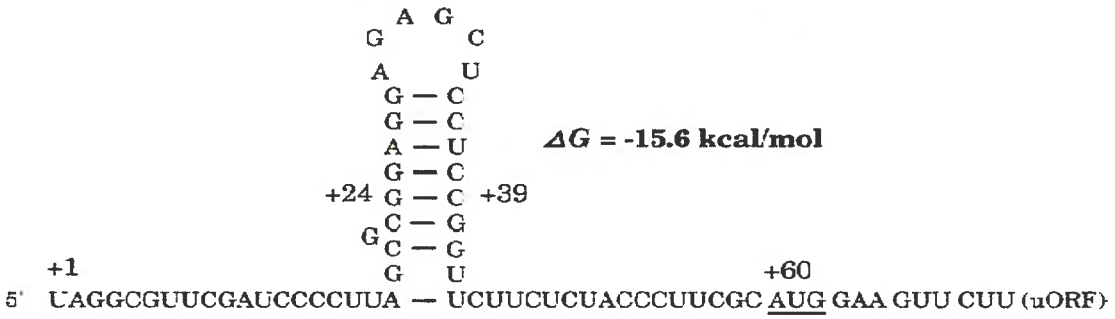

Figure 3. A potential hairpin structure in the 5' leader of Lc mRNA. The ΔG value of the hairpin was calculated according to Tinoco et al. (1973). Lc mRNA is numbered according to Ludwig et al. (1989) with the uORF initiation codon underlined. A $G_{24} \rightarrow C$ substitution reduces the ΔG value from -15.6 to -5.4 kcal/mol, and a subsequent $C_{39} \rightarrow G$ change restores the ΔG value. Picture taken from Wang, L. and S. R. Wessler (2001). "Role of mRNA secondary structure in translational repression of the maize transcriptional activator Lc (1, 2)." Plant Physiol **125**(3): 1380-7.

hairpin is important for repression because its destabilization increases translation of the upstream open reading frame (ORF) and the downstream ORF (Wang and Wessler 2001). However, translation of the upstream ORF is not required for the hairpin-mediated repression.

Patterns in the Genetic Code

In 1972 White et al. suggested that the redundancy in the genetic code permits extensive variation of the nucleotide sequence and allows the satisfaction of requirements for both protein structure and RNA structure (White, Laux et al. 1972). Ball proposed three alternative hypotheses on the relationship between amino acid sequences and

mRNA secondary structure. First, the choice of codons and their sequence in the message could be independent of the resulting secondary structure of the mRNA. Second, optimization of mRNA secondary structure may occur only within the limits of encoded amino acid sequence. Third, selection pressure for specific RNA secondary structure could affect the choice of nucleotide at both synonymous (position of anti-codon which recognizes more than one base in the codon) and non-synonymous (position of anti-codon that does not allow swapping of multiple bases in the codon) positions. In 1974 Fitch examined these hypotheses and found evidence of the use of the degeneracy of the genetic code to optimize base pairing in mRNA molecules. He discussed the third hypothesis as being biologically plausible although he did not find evidence for or against the notion that the needs of RNA structure and function must compete with the needs of protein structure and function.

Virtually all mRNA sequences carry a 3-base periodical pattern (wobble position), presumably involved in the translation frame monitoring mechanism. It was reported that periodical patterns complementary to the proof-reading site in the ribosome and presumably involved in the translation frame monitoring mechanism have been found in many transcripts (Lagunez-Otero and Trifonov 1992). It was shown that synonymous substitutions affect mRNA translation in different organisms (Ikemura 1985). Strong mRNA secondary structures formed due to gene-specific codon usage have been implicated in discontinuous translation and pauses in synthesis of insect silk fibroin, chicken collagen and other proteins (Mita, Ichimura et al. 1988). These and similar works gave rise to the expectations that secondary structures can interfere with translation and therefore should be avoided in mRNA coding regions. Contrary to this opinion,

significant biases in favor of local RNA structures have been found in several bacterial species and the yeast (Katz and Burge 2003). Although evolutionarily conserved local secondary structures were described in eukaryotic and mammalian mRNAs and pre-mRNAs (Meyer and Miklos 2005), no conclusive evidence has been found to confirm or disprove the hypothesis that selection for RNA structure can lead to nonoptimal amino acid usage. Correlations between mRNA and protein secondary structures have been noted (Luo, Jia et al. 2004). A phenomenological model on the relation between structure preference and translational efficiency or accuracy was proposed. It is pointed out that the structure preference of codons is related to the distribution of mRNA stem/loop content in three t-RNA copy number(TCN) regions (Luo, Jia et al. 2004). Seffens and Digby reported that native mRNAs have a lower calculated folding free energy than random sequences. An examination of 51 mRNA sequences in GenBank revealed that calculated mRNA folding is more stable than expected by chance. Seffens and Digby, noticed that free energy minimization calculations of native mRNA sequences are more negative than randomized mRNA sequences with the same base composition and length. Randomization of coding region of genes yields folding free energies of less magnitude than the original native mRNA sequence. Randomization of codon choice, while preserving original base composition, also results in less stable mRNAs. They concluded that a bias in the selection of codons favors the potential formation of mRNA structures which contribute to folding stability.

Neural networks/ Backtranslate sequences/Pattern detection

An artificial neural network (NN) is an information processing paradigm that is inspired by the way the densely packed and interwoven neurons of the brain relay

information (Abdi, 1994). It is a mathematical model or a set of algorithms that emulate information processing properties of the brain. NN programs have large numbers of interconnections that are linked to connection weights that may be thought of as synaptic connections. Neural networks have been used to examine sequence patterns identifying coding regions in genomic DNA (Snyder and Stormo, 1993).

A small NN was trained on amino and nucleic acid sequences to test the NN's ability to predict the correct codon given only an amino acid sequence (White and Seffens, 1998). Different network configurations were used with varying numbers of input neurons that represented amino acids and a constant representation for the nucleic acid. A multi-layer backpropagation network of one hidden layer with 5 to 9 neurons was used. In the best-trained network, 93% of the overall bases, 85% of the degenerate bases, and 100% of the fixed bases were correctly predicted. The training set was composed of up to 60 human sequences in a window of up to 25 codons at the coding sequence start site. Different input configurations for amino acid representations were designed and evaluated. It was found that the input configuration was not important for NN performance, so a simple unitary representation of amino acids is adequate.

In previous work conducted by Pratt 2003, the NN configuration 2LM was trained with two human data sets of negative Z-scores and of positive Z-scores, which produced different results. Java NNS (Zell A., 1995) was used allowing amino acid window sizes of at least 20. This window size was important to encompass typical stem-loop structures that would be found in mRNA. The configuration 2LM trained on the data set with positive Z-scores for mRNA secondary structure produced a total testing accuracy of 93%. However the network 2LM trained with the data set for high mRNA secondary

structural bias gave a significant increase to 97.5% total testing accuracy and an increase to 92% in degenerate base predictive success. These numbers were significantly better than previous results obtained from White and Seffens (1998) where the predictive accuracy for degenerate bases was 80% (Training set 60S-20c) and 85% (Training set 60S-10c) for a similar NN configuration. These results furnish evidence which support the idea that mRNA secondary structure increases NN learning and predictive capabilities (Pratt 2003).

Further, those studies support the idea that longer training or larger data sets yield better backtranslation accuracy. Pratt performed additional analysis of the predicted codons to determine whether or not the predictions changed at different positions within a sequence for the same amino acid. Interestingly, they found that many of the alternate codons were being predicted correctly for degenerate amino acids. In conclusion, this suggests that the NN is not necessarily using the most frequently used codon, but is learning patterns within the data set which allows the NN to predict which codon to use in relation to neighboring coded amino acids (Pratt 2003). Mandy Lucas in 2006 took the codon analysis one step further by investigating global mRNA secondary structure and local codon choice in human genes. Lucas observed that human sequences partitioned based on their secondary structure gave lower sum squared errors for the directory characterized with excess secondary structure trained by neural networks; suggesting that there are more detectable sequence patterns in genes with more secondary structure than in genes exhibiting the least amount of secondary structure.

Folding Free Energy

Free Energy is the energy thermodynamically available to do work. For example, the energy required for a protein to fold properly is also known as free energy. Normally a system is believed to be in its most stable conformation when minimum amount of free energy is exerted. When many homologous RNA sequences are available, the standard technique for determining the secondary structure is comparative sequence analysis (James et al., 1989; Pace et al., 1999; Woese et al., 1983). When there is only one or a few known sequences for RNA, free energy minimization can also be used to predict secondary structure models that can be tested against experimental data such as chemical modification and site directed mutagenesis (Walter et al., 1994a; Hofacker et al., 1994; Jaeger et al., 1989; Mathews et al., 1997). Thermodynamic parameters for the prediction of free energy of folding are at the heart of algorithms for secondary structure prediction (Zuker, 1989; McCaskill, 1990; van Batenburg et al., 1995; Gultyaev et al., 1995). Parameters based on a nearest-neighbor model (Xia et al., 1998) are well determined experimentally for Watson-Crick pairs, but helical regions are not that determined. The remaining nucleotides are in unpaired regions, mostly loops. Recent studies have demonstrated that the stabilities of loops are highly sequence dependent (Schroeder et al., 1996; Serra et al., 1997; Wu et al., 1995; Xia et al., 1997).

As previously mentioned, traditional methods for calculating DNA and RNA secondary structure have mostly used free energy minimization for single sequences and phylogenetic comparisons for homologous, alignable RNA's that use compensatory mutations as evidence for conserved base pairs (Zuker 2000). Also It has been proposed that RNA sequences can be classified according to whether they are more or less stable

(thermodynamically favorable or not) in calculated folding free energy as compared to randomized sequences (Seffens, 1999). In 2005, Dr. Adam Davis reported with a larger set of human genes that the native genes were more stable as compared to randomize (shuffled nucleotides of native sequence to create 10 sequences) sequences (Davis 2005). mRNA secondary structures that contribute to calculated folding free energies may be involved in gene regulation mechanisms, intron splicing, or steady state mRNA levels. When calculating energy minimization, free energies are assigned to base pair stacks and to loops, and are summed to calculate the overall free energy difference of folding (Zuker 2000). Base pair stacks take into account both hydrogen bonds and stacking effects. Loop energies comprised an entropic term for loss of conformational freedom and other terms that take into account mismatched pair stacking, co-axial helix stacking, single base stacking and empirically derived corrections (Zuker 2000). Such energies are referred to as 'nearest neighbor rules', meaning in helices individual terms depend not on single base pairs but on base pairs that are conditional on their adjacent pairs.

Z-scores

Z-score is a statistical method of standardizing data on one scale so a comparison between variables will remain consistent. The Z-score value indicates how many standard deviations an observation is above or below a mean (Larson and Marx 2000). There are various algorithms for calculating Z-scores based on statistical distribution assumptions. The Z-score calculations used by Workman and Krogh were not clearly defined (Workman and Krogh 1999). The Z-score calculated by Katz and Burge used a standard normal distribution (Katz and Burge 2003). The appropriate Z-score equation is vital in determining how far a sample of interest is from the mean of interest represented by zero.

Different Z-score methods are used for different data sets. A problem associated with determining significant differences of an experimental parameter such as folding free energy between mRNAs in a data set is the blocking of the variation between mRNAs (Davis 2005).

The Z-score is the preferred method for analyzing the difference between two means or a sample and a mean (Montgomery 2005). However, it is common in biological research to find the use of a Z-score that is not the most accurate for a particular research design (Seffens and Digby, 1999, Katz and Burge, 2003, Workman and Krogh 1999). Seffens and Digby, 1999, Katz and Burge, 2003, Workman and Krogh 1999 all published on the significant differences between the native folding free energy and its randomly shuffled folding free energy of the coding regions of mRNAs. All used the Z-score with the standard deviation as the denominator, known as the '*Standard Normal Distribution*' (SND). This Z-score is used to compare two samples, and should not be used to compare a sample to a mean. This is because of the theories of the '*Central limit Theorem*' (Montgomery, 2005). Because most Z-score differences are in the denominator, if the SND Z-score is used the sample data Z-score will not be as far away from zero. The appropriate Z-score to use when comparing a sample to a mean is the *'Normal distribution'* (ND). In experiments where the ND Z-score is 1.64, if SND Z-score was used, it would result in a number less than 1.64 say 1.05. Using the SND Z-score will reduce the level of significant difference between test samples and its mean. It is not an appropriate method to use to determine the distance a sample is from a mean. Using the ND Z-score results in an increase in the number of sequences where the native

folding free energy is at least 1.64 standard deviations away from its randomized folded free energy mean.

Purpose of research

Free energy minimization calculations of native mRNA sequences have been reported as more negative than randomized mRNA sequences of the same nucleotide composition (Seffens and Digby 1999). This suggests a possible bias in codon choice favoring mRNA structures that have greater folding stability, as proven by greater folding free energies. mRNA secondary structures that contribute to calculated folding free energies may be involved in gene regulation mechanisms, intron splicing, or steady state mRNA levels (Sandberg and Mulroney, 2001). Structural elements of mRNA are known to play integral roles in mechanisms regulating translation and mRNA stability. Since message turnover is an important component of gene regulation, it is not surprising to find that message stability characteristics of key growth regulatory genes are tightly controlled as a group (Davis 2005).

Since there have been numerous reports linking secondary structure to regulation of gene expression we wish to explore the inherent properties of various transcriptomes. What factors influence stable folding free energies reported by Seffens and Digby? The genetic code may be optimized in part to form stable mRNAs to protect them from degradation (Digby and Seffens, 1999). This may be a relic from the postulated RNA world before DNA became the repository of genetic information. Does gene length affect the secondary structure of a sequence? Are there any patterns that exist which allows for one organism to express excess or minimal secondary structure? All of these

questions are the basis for which this research leans on because the answers from these questions will allow for ways to possibly link diseases to folding free energies. Secondary structure and folding free energies could aid in creating gene profiles of various disease associated genes. Possible correlations between folding free energies and gene expression could significantly enhance understanding of cancerous versus non-cancerous gene types.

CHAPTER 3

MATERIALS & METHODS

Compiling Transcriptomes

Previous studies (Seffens and Digby, 1999, and Workman and Krogh, 1999) investigated a dataset of only 50 mRNA sequences. In studies done by Dr. Adam Davis in 2005, he investigated the transcriptomes of human, mouse, and Arabidopsis (plant) (Davis 2005). We wish to expand the scope of transcriptomes because this will help to elucidate secondary structure significance influenced by dinucleotide compositions in genes. The relationship between mRNA basepair-frames and mRNA secondary structure folding free energies of *Homo sapiens* (human), *Mus musculus* (mouse), *Gallus gallus* (chicken), *Pan troglodytes* (chimpanzee), zebrafish, *Trypanosma brucei*, *Strongylocentrotus purpuratus*, *Theileria parva* (protozoa), and *Cryptococcus neoformans* (fungi), *Apis mellifera* (honeybee), *Drosophilia melanogaster* and *Arabidopsis* mRNA will be analyzed. The transcriptomes will be downloaded from the National Center for Biotechnology Information (NCBI) website and then folded. In addition we will generate randomly shuffled versions of each file and fold those sequences. All of this data will be gathered using a program called Rgather for the generation of the Z-scores. The more transcripts available the greater statistical

secondary structure significance influenced by mono and dinucleotide composition in genes. The Monoshuffle program preserves the same mononucleotide composition and length as native mRNA. These sequences have the same number of A, C, G, and Ts; whereas the Dishuffle program additionally preserves the dinucleotide compositions. These sequences have the same number of AA, AC, AG, AT, CA …etc compositions.

Multi-Regression Analysis

The second aim is to perform a multi-regression analysis on the base content of the particular genomes. This analysis will be performed in an attempt to explain the trend of the negative values characterized as genes with excess secondary structure being compiled of the various transcriptomes. Multiple Regression studies the relationship between several independent and dependent variables. The variables that we wish to relate are the base content with the level of complexity in a certain genome. Also we wish to study the comparison between base compositions of the genome with its relative Z-score value. For example, determining correlations between dinucleotide and trinucleotide content with secondary structure as computed by the Z-score. Comparing species based on normalized gene length and deciphering if folding free energies are affected by the length of the sequence. Transcriptome comparative studies will be conducted to determine if base composition bias affects the folding properties of genes. This will clarify if base composition, whether mononucleotide or dinucleotide, is a selective property of mRNA or if it is impacted largely by nonselective factors such as DNA mutation and repair processes. To answer these questions statistical techniques will be applied to perform appropriate regression analysis of folding energies and z-scores. This will identify factors that have the greatest impact on Z-scores in the transcriptome.

The goal of this study would be to see if the base compositions of a particular organism have an influence on the Z-score.

Artificial Neural Network

Using artificial neural networks we wish to examine sequence patterns throughout the different transcriptomes. It is disputed that the genetic codes available in nature are arranged in such a way that they are resistant to errors (Marquez, 2005). Since there is evidence that native sequences are thermodynamically lower in folding free energy, the question here is, what is causing this trend? What are the evolutionary pressures contributing to this pattern? To do this we will analyze the genes of several species and distinguish their sequence pattern based on pattern recognition from artificial neural networks. We will partition genes based on secondary structure as computed by Z-scores to determine if the artificial neural network learns at different rates. We predict that the transcriptomes with more secondary structure will have fewer errors in the network. The neural network should have a lower error rate with transcriptomes that display very negative z-scores and/or excess secondary structure.

Computer Equipment

Analysis in this investigation was done on twenty-two (22) Intel x86-based Windows 2000 PC workstations. These workstations were linked together as a Network of Workstations (NOW). A Dell Power Edge 4600 Intel Xeon Windows 2000 dual processor server served to control the NOW workstations and is used for file storage. All statistical analysis was carried out using Microsoft Excel, SPSS statistical software

package, and sequence folding using RNAStructure (v4.2) based on MFold program (Zuker M., 1999).

Identification of genomes sets

Transcriptomes were obtained from the Reference Sequence Project (RefSeq) at NCBI (http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/ handbook/ch18d1.pdf). Each RefSeq RNA sequence represents a distinct transcript produced from a particular gene representing a gene model. All gene models based on a particular RefSeq RNA are compared, and the best one is selected. Extra models representing paralogs are included with the mRNA- and EST-based models. Between builds, RefSeq RNAs are refined based on a review of related gene models and transcript alignments produced during the genome annotation process. Human mRNA sequences were extracted from a reference set called "rna.gbk.gz" obtained from NCBI

(ftp://ftp.ncbi.nih.gov/genomes/Hsapiens/RNA /rna.gbk.gz) dated March 10, 2004, Version: 3. Mouse mRNA sequences were extracted from a reference set called "rna.gbk.gz" obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/ M_musculus /RNA/rna.gbk.gz) dated March 31, 2003. Arabidopsis mRNA sequences were extracted from a reference set called plant1.rna.gbff.gz, plant2.rna.gbff.gz obtained from NCBI (ftp://ftp.ncbi.nih.gov/refseq/release/plant/ (names of file)) dated May 6, 2004. The other transcriptomes were extracted from Reference Sequence Project as well in same fashion as the others species that I mentioned earlier. At NCBI's refseq database the transcriptomes were stored as compressed zip files with GBFF file extentions and to retrieve the necessary RNA files we used a program called "SplitRNA.exe" which

separated each mRNA. The mRNA sequences were split into the following categories, mRNA with sequence lengths larger than 3,000 bp and have ambiguities (*Lgambig2*), mRNA that were larger than 100 bp and less than 10,000 bp in length (*Not2Big2*), mRNA that were smaller than 100 bp with ambiguities (*Smambig2*), mRNA that still contained their introns "Pre-RNA" (special2) and genes that were larger than 10000 bp (*Too_big*) directories. For much of the data presented here the Not2big directory was chosen because this category was comprised of mainly of coding region (CDS) and in addition to the CDS we did not want sequences that were really huge as this made the folding time very long.

mRNA Randomization Procedure

The genes from the *Not2Big2* directory on the NOW server were split into directories totaling 1000 genes each. The NOW server is networked to a directory call "Folding" on each of the twenty-two workstations. From the server, each directory containing 1000 mRNA sequences was loaded onto each workstation folding directory. All mRNA sequences were shuffled using two methods.

1. Monoshuffle Program-- Preserves the same mononucleotide composition and length as native mRNA. These sequences have the same number of A, C, G, and Ts.
2. Dishuffle Program-- Similar to Monoshuffle except it additionally preserves the dinucleotide compositions. These sequences have the same number of AA, AC, AG, AT, CA …etc compositions. Each gene was shuffled ten times by the "Monoshuffle" and/or the "Dishuffle" programs.

Calculating the simulated mRNA secondary structure

In order to calculate the secondary structure throughout the transcriptomes mRNA files had to be processed and formatted properly. Once the mRNA sequences were shuffled, the shuffling program created a "$$file$$.bat" file. The $$file$$.bat file was used to execute RNAStructure for folding. This bat file is started on each of the NOW workstations after the shuffling program has shuffled all of the genes in the local Folding directory. A program named "Rgather" collects the results and calculates statistical data for each mRNA (figure 4), from RNAStructure program output. The data output file (*.1LM) from the Rgather program is in simple "txt" format. This was loaded into Microsoft Excel Program and formatted as a comma delimited (CSV) file. All the statistical information was imported into a Microsoft Excel spreadsheet.

Z-scores

Previous studies have used Z-scores to determine significant levels of differences between native genes and shuffled versions across sets of genes. The *Z*-score (Le and Maizel 1989) is the number of standard deviations by which the minimum free energy (MFE) of **x** deviates from the mean.

Statistical definition for Zone of Acceptance (Z-score)

- To test individual samples or in this case genes, the statistical hypothesis is:
- To test how far the native folding free energy is from the mean.

Ho: $\mu_N = \mu_R$ (No Difference)

Ha: $\mu_N \neq \mu_R$

The null hypothesis is that there is no difference between the native and random folding free energy.

Normal distribution: $Z^* = \frac{\mu_N - y}{\sigma / \sqrt{n}}$

Where y = Mean of the 10 random folding free energies

μ_N = Native folded free energy

σ = Standard deviation of the average randomized folded free energy

$\sqrt{n}$ = Square root of the total number of genes

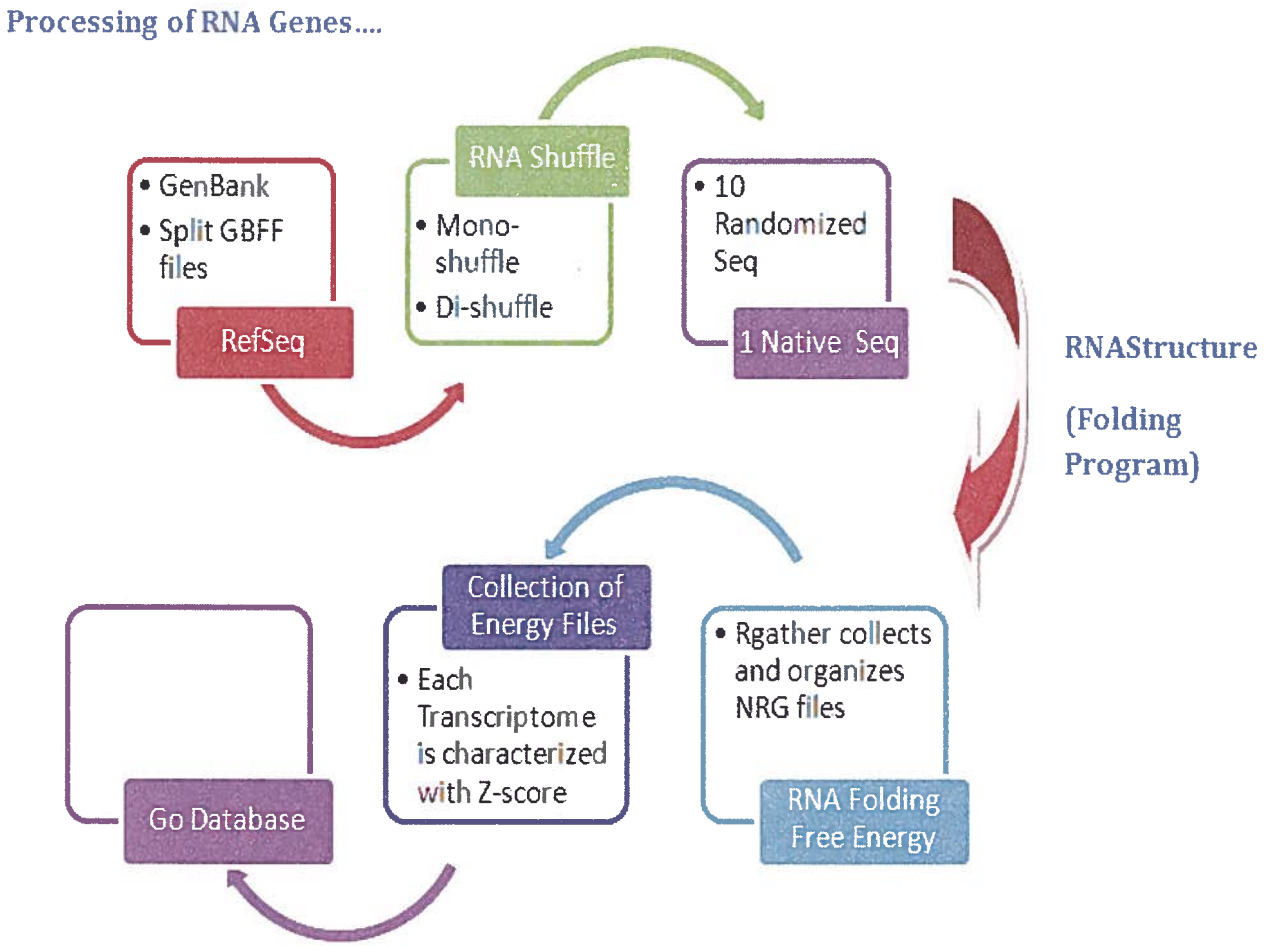

Figure 4. The process by which RNA files are folded and calculated. The sequences used for this project was gathered from the Reference Sequence Project and folded under the dishuffle method. Upon folding completion the sequences was gathered so that the Z-score calculation could take place. Z-scores were measures used to characterized the transcriptomes based on secondary structure.

Neural network pattern analysis

Sequence patterns have been examined by others using neural networks to identify coding regions in genomic DNA (Snyder and Stormo, 1993), for detecting errors in mRNA splice sites (Brunak et.al., 1990; Ogura *et al.*, 1997), for predicting the mechanism of action of cancer drugs (Weinstein, et.al., 1992), and for predicting the secondary structure of proteins (Qian and Sejnowski, 1988; Holley and Karplus, 1989; Bohr *et al.*, 1988). The performance of several neural networks in identifying coding

regions in genomic DNA sequences was evaluated (Snyder and Stormo, 1993; Kraemer, et.al., 2001). A backpropagation NN was used, and the highest accuracy was obtained with a NN called GeneParser. It predicted 75% of the exons correctly as exons. The trained network can then be used to identify genes in GENBANK. There were several biological parameters involved in designing this network. Among them were codon usage, informationally rich regions, length distribution among introns and exons, mutual information, and presence of donor and acceptor sites. In noncoding regions of the genome there are large amounts of repetitive DNA sequences; however, coding regions tend to be informationally rich. Both intron and exons have characteristic length distributions that can be used to classify them (Kraemer et.al. 2001).

Binary Encoding

For training purposes, it was necessary to provide the computer generated neural network architectures with a way to understand the information that it was attempting to learn. This was because the neural network does not allow for direct representation of nucleic or amino acid sequences. Thus the simplest way to do this was to encode the training set (mRNA sequences) into a binary form. Consequently, a numeric string of ones (1) and zeros (0) in binary code were used to depict each of the twenty amino acids and subsequently each nucleic acid sequence. For example the alphabetically first amino acid Alanine, is encoded by using a one and nineteen zeros (10000000000000000000). Subsequently, the position of the one would shift to the right dependant on the single letter abbreviation of the amino acid alphabetically. This means that the amino acid Cysteine "C" would also be represented by a one and nineteen zero, but the position of the one here would be shifted one place to the right (01000000000000000000).

Software (Java NNS)

Most of the analysis in this investigation was done on five (5) Core 2 Duo processor computers. Java Neural Network Simulator (JavaNNS) was the simulator of neural networks used for this project and is the successor of Stuttgart Neural Network Simulator (SNNS). It is based on its computing kernel. The simulator kernel operates on the internal network data structures of the neural nets and performs all operations of learning and recall. It can also be used without the other parts as a C program embedded in custom applications. It supports arbitrary network topologies and supports the concept of sites. SNNS can be extended by the user with user defined activation functions, output functions, site functions and learning procedures, which are written as simple C programs and linked to the simulator kernel. Three auxiliary programs written in "C" programming language were also utilized; PatternMaker.exe, pattern_list.exe, and pattern_test.exe. These programs prepared sequences for the training and validation sets. PatternMaker.exe program formats the mRNA files so that they could be fed into the neural network. Pattern_list.exe simply lists all of the mRNA files that are fed into the neural network. It becomes extremely useful when an error occurs in the network and a file in missing, pattern_list.exe is great because it allows deciphering of the missing file by matching the input file with the output. Pattern_test.exe ensures that the input file, a pat.exe file is 'error-free' to permit successful neural network learning.

Training Neural Networks

When preparing training for pattern detection in species using artificial neural networks there were several steps executed to ensure consistency for the data. First, sequence collection was done by extracting the sequences based on their secondary

structure characterized by the Z-score. Three sets or directories of Z-scores were partitioned into Negative-Z, Near Zero-Z, and Positive-Z. However, before the RNA files were imported into the neural network, the files for each transcriptome had to be sorted and separated into specific sets (figure 5). This process was carried out by gathering the data sheets of each transcriptome and dividing the files based on the Z-scores. All the organisms used for this research had at least five thousand RNA files, therefore using one thousand RNAs for each directory was more than sufficient in terms of sample size. The negative-Z sets represented the Z-scores that expressed the most secondary structure in RNAs (showed the most negative Z-score values). In contrast, the set labeled positive-Z symbolized the RNAs that expressed the least amount of secondary structure or the positive Z-score values. The near zero-Z set corresponds to the Z-score values that have mixture of both positive and negative Z-scores. In each of the directories a program called "Pattern Maker.exe" formatted the sequences into training sets compatible for the neural network program. The neural network (JavaNNS) is trained using a large set of specific mRNAs, in this case approximately 1,000 and from there the sum squared error rates were recorded. The neural network was operated under random weights with parameters of-1.0 and 1.0.

The training sets were allowed to run up to 100,000 generations. One generation is one complete cycle through a training set. So after every gene or sequence in a training set is learned by the neural network one generation is completed. Parameters were designated to specify the amount of neural network generations. Within the java neural network there's a Log that states the error at the specified generation. Under normal

conditions, as generations increase the error decreases because over generations the neural networks learns the patterns of the file.

Neural network learning parameters were assessed to determine the amount of influence on neural network performance. Sum square error logs recorded the learning error at specified generations. As neural network generations increased, overall the learning error decreased. Recent studies of the problem of training and generalization in neural networks have suggested (Grossberg and Levine, et al. 1987) that a critical number of examples exist, above which the generalization error falls off exponentially fast, due to a gap in the spectrum of generalization error (**ε**). In contrast, in the present work such a critical number does not exist. Instead, whenever a large number of genes can be learnt, **ε** approaches a power log. The power law behavior of **ε** is a manifestation of the absence of a gap in the spectrum of **ε**.

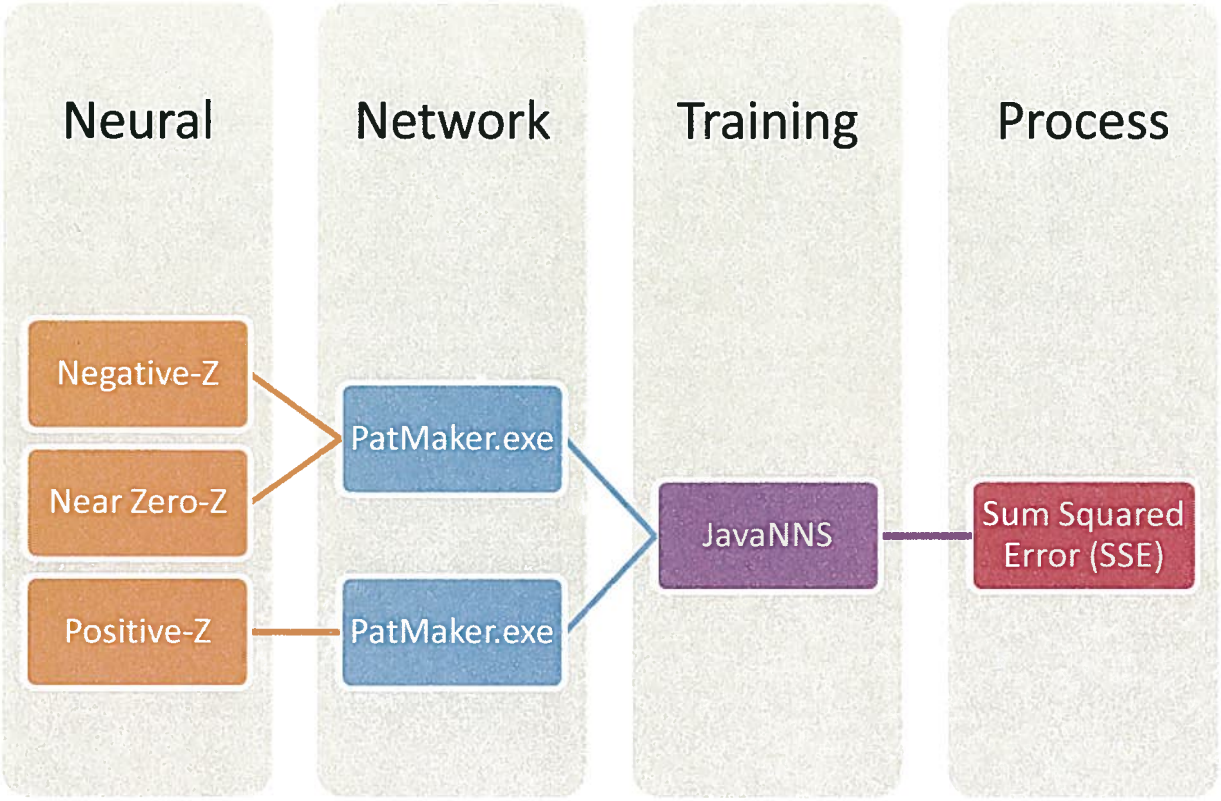

Figure 5. Process of training the neural network

CHAPTER 4

RESULTS

Expansion of Data Sets

We set out to examine the statistical difference between secondary structure folding free energies of the native and randomized mRNA sequences across species. Dr. Adam Davis expanded the data for human genes from 50 Monoshuffled mRNAs (Seffens and Digby, 1999 and Workman and Krogh 1999) to 6551 mRNA for the Monoshuffled method and 6221 mRNA for the Dishuffled method. Also Dr. Adam Davis developed the data sets for the *Arabidopsis* (plant) and the *Mus musculus* (house mouse). The *Gallus gallus* (chicken), *Pan troglodytes* (chimpanzee), *Stronglocentrotus purpuratus, Trypanosoma brucei*, *Danio rerio* (zebrafish), *Drosophilia melanogaster* (fruit fly), *Apis mellifera* (honeybee), *Theileria parva* (protozoa), and *Cryptococcus neoformans* (fungi) were all transcriptomes that we compiled and expanded. These data sets were expanded and developed throughout my tenure as a graduate student. Free energies of folding associated with secondary structure were calculated for all the transcriptomes using RNAstructure software (Zuker, *et.al.*, 1999). mRNA sequences were also shuffled under a constant dinucleotide constraint and folding energies calculated for 10 shuffled sequences for each mRNA. A Z-score was then calculated for each mRNA as: (Native-Avg. Shuffled)/ (Std. Dev. of shuffled set) = z-score, (Le and Maizel 1989). Single-value means of each transcriptome were calculated as an average of all z-scores. In the cross species comparative study between the human, chimpanzee,

mouse, *Arabidopsis* (Plant), *Danio rerio* (zebrafish), *Theileria parva, Cryptococcus neoformans, Drosophilia melanogaster* , *Strongylocentrotus p., Apis mellifera, Trypanosoma brucei*, and chicken transcriptomes, it was found that the chicken transcriptome had a greater proportion of stable mRNAs than human, and the mouse has more stable mRNAs than the plant (table 1). The procedure for calculating the most stable mRNAs was done by counting the number of genes that possesses a negative z-score and then dividing that number by the total number genes in that transcriptome. For example, out of 6551 human mRNAs, 80% (5, 214 genes) of the native free energies were more negative than their randomized sets (table.1).

Table 1 All of the transcriptomes compiled using RNAStructure.

Transcriptomes Compiled	**Z-Score**	**% of Neg. Native genes**	**# of Sequences**	**Ver. of RNAStructure**
Chicken	-2.45	94%	5141	4.2
Chimpanzee	-1.94	91%	606	4.2
Human	-1.698	80%	6551	3.7
Strongylocentrotus	-1.07	74%	1069	3.7
Honeybee	-0.895	71%	1699	4.2
Trypanosoma	-0.405	59%	3208	3.7
Mouse	-0.253	56%	11967	3.7
Plant	-0.197	55%	14594	3.7
T. Parva (protozoa)	-0.245	55%	523	4.2
Zebrafish	-0.3817	54%	6318	4.2
Drosophila melanogaster (Fly)	-0.09162	51%	12536	4.2
Cryptococcus Neo (fungi)	0.71	28%	3924	4.2

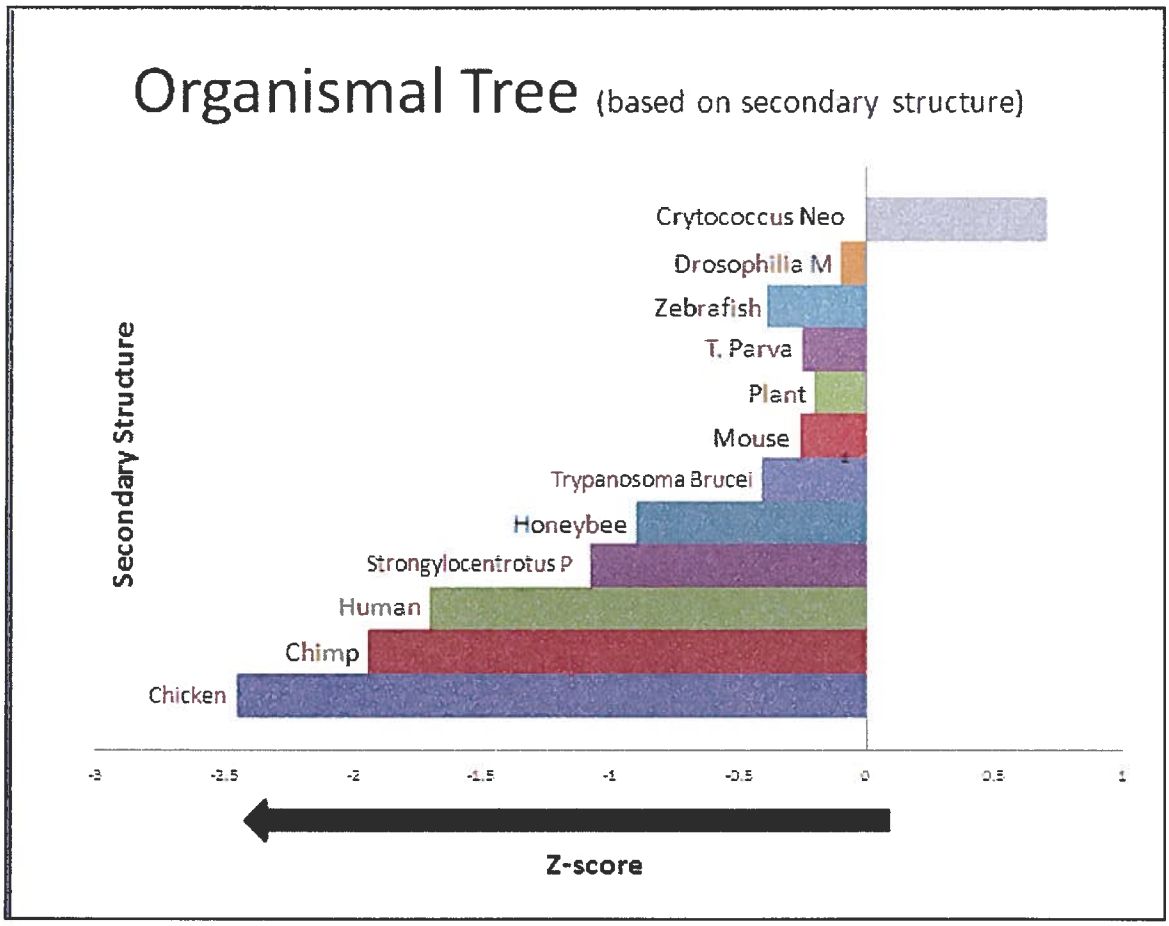

Figure 6. Transcriptomes compiled and ranked based on secondary structure characterized by Z-scores.

Determining sufficient sample size

When comparing transcriptomes, it is statistically best to use the same number of sequences for each species. Initially, we were able to get several transcriptomes to match up in terms of number of sequences; but as more transcriptomes were extracted from NCBI many of the zip files contained errors which reduced the number of sequences. So this dilemma forced us to conduct an experiment to determine the number of sequences needed to represent a full transcriptome. In other words, what number of genes is required for a confident calculation of the average Z-score for the transcriptome? We accomplished the task of figuring out how many sequences are needed and we did this by randomly selecting genes along with their reported Z-score from the human transcriptome. We created ten groups of different sample sizes with randomly selected genes and their Z-scores. After calculating the average for each group, we plotted the average of the group with the number of sequences. We analyzed the averaged Z-score for groups that were comprised of 10, 20, 40, 50, 75, 100, 150, 175, 200, 250 number of sequences. What we observed was that above one-hundred sequences the variability in the average Z-score began to decrease. Between ten and seventy-five sequences the variability increased as denoted by the spread (figure 7). We performed this study on mouse and chimpanzee transcriptomes as well both reporting the same results (figure 7). We concluded from this information that one-hundred sequences are sufficient enough to characterize a transcriptome.

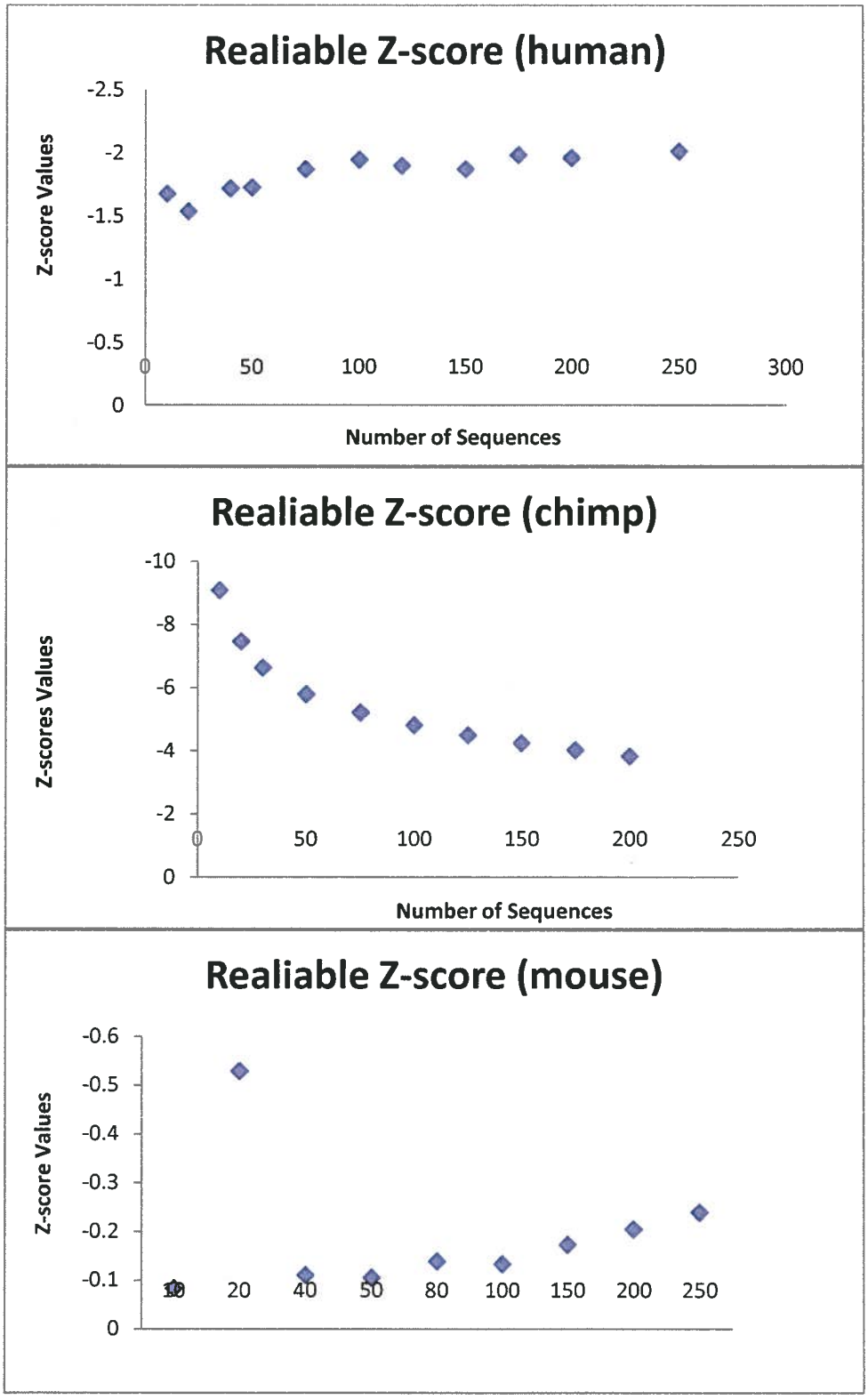

Figure 7. Number of sequences required to characterize a transcriptome. Human, Chimpanzee, and Mouse transcriptomes were used. One-hundred sequences seem to be a sufficient sample size to compare amongst transcriptomes.

Number of Shuffles

The number of shuffles required to randomize a sequence to calculate gene Z-scores needed to be determined. Seffens and Digby in 1999 considered that ten (10) shuffles would be adequate to randomize a gene. But after conducting a trial run with human genes we found out that ten may not be the correct number of shuffles. We randomly chose ten human genes using the RAND function in Microsoft excel and shuffled each of them one-hundred times. The ten genes chosen represented a variety of different processes such as translation (ribosomal protein S29, NM_001032), transcription (zinc-finger protein124, NM_003431), and cellular movement (cadherin15, NM_004933). After shuffling (randomizing), we folded each of the sequences using RNAstructure to compute the folding free energies. We took the average folding free energy values at several different shuffle counts (10, 20, 50, 70, 80, and 100) to find out what the variability was (figure 8). So out of ten different genes eight of the genes had plots that displayed the variability decreasing once the shuffle number reached 10. Five percent error is allowed in statistics assuming the sample size, n is not large. All the genes reported similar variation trends depicted by the error bars. From this data we concluded that ten shuffles were statistically sufficient to represent a randomized sequence. Also, when performing the same experiment using the mouse transcriptome ten shuffles was sufficient as well when adding the error bars. Also, with RNAStructure only being responsible for processing ten variations of genes as oppose to fifty, this made the computing time much quicker.

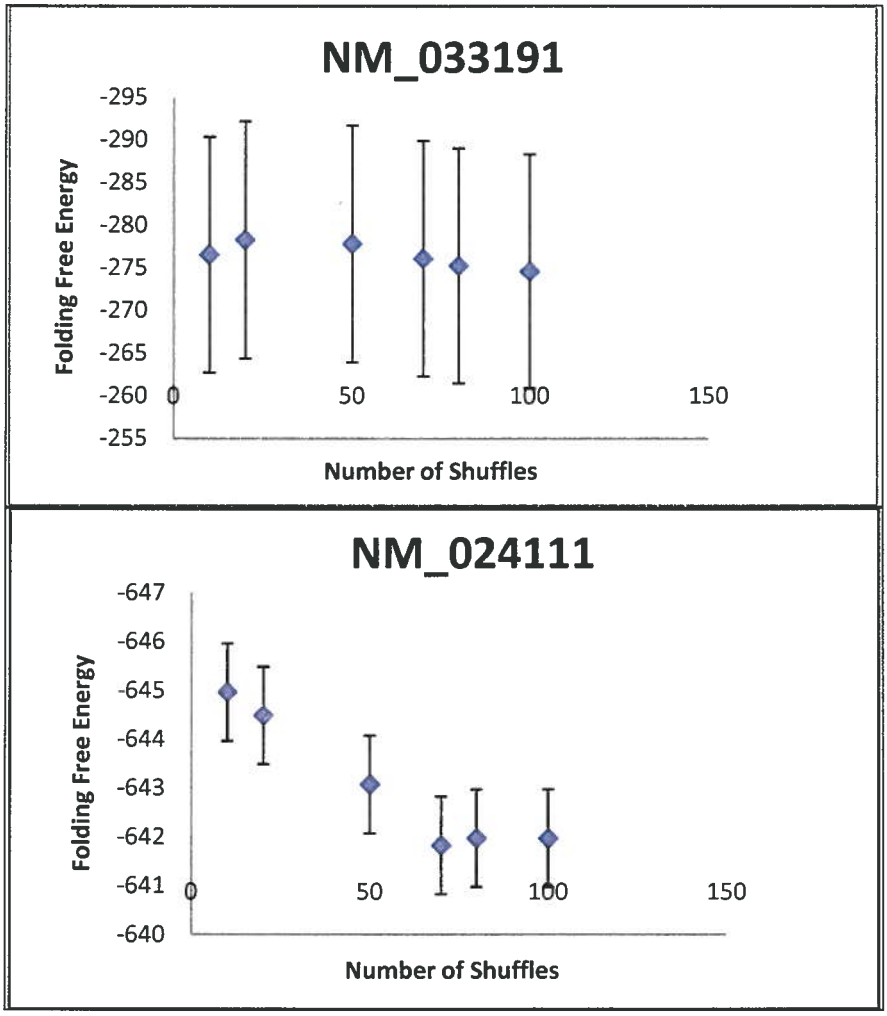

Figure 8. Initially 50 shuffles were deemed to the correct number of shuffles needed.

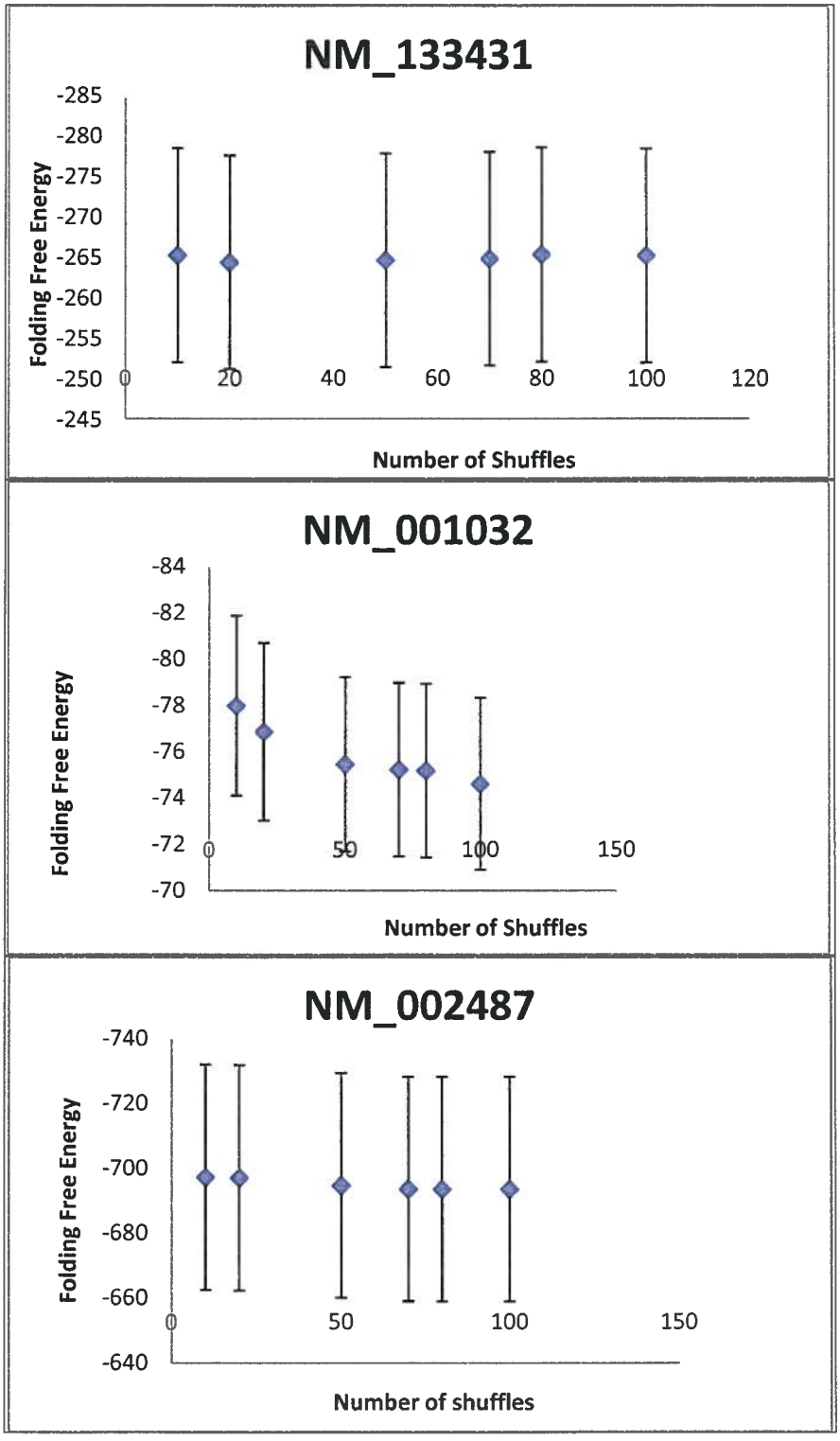

Figure 8. The error bars denotes that ten (10) shuffles are enough to adequately randomize a sequence as there isn't much variation after ten shuffles.

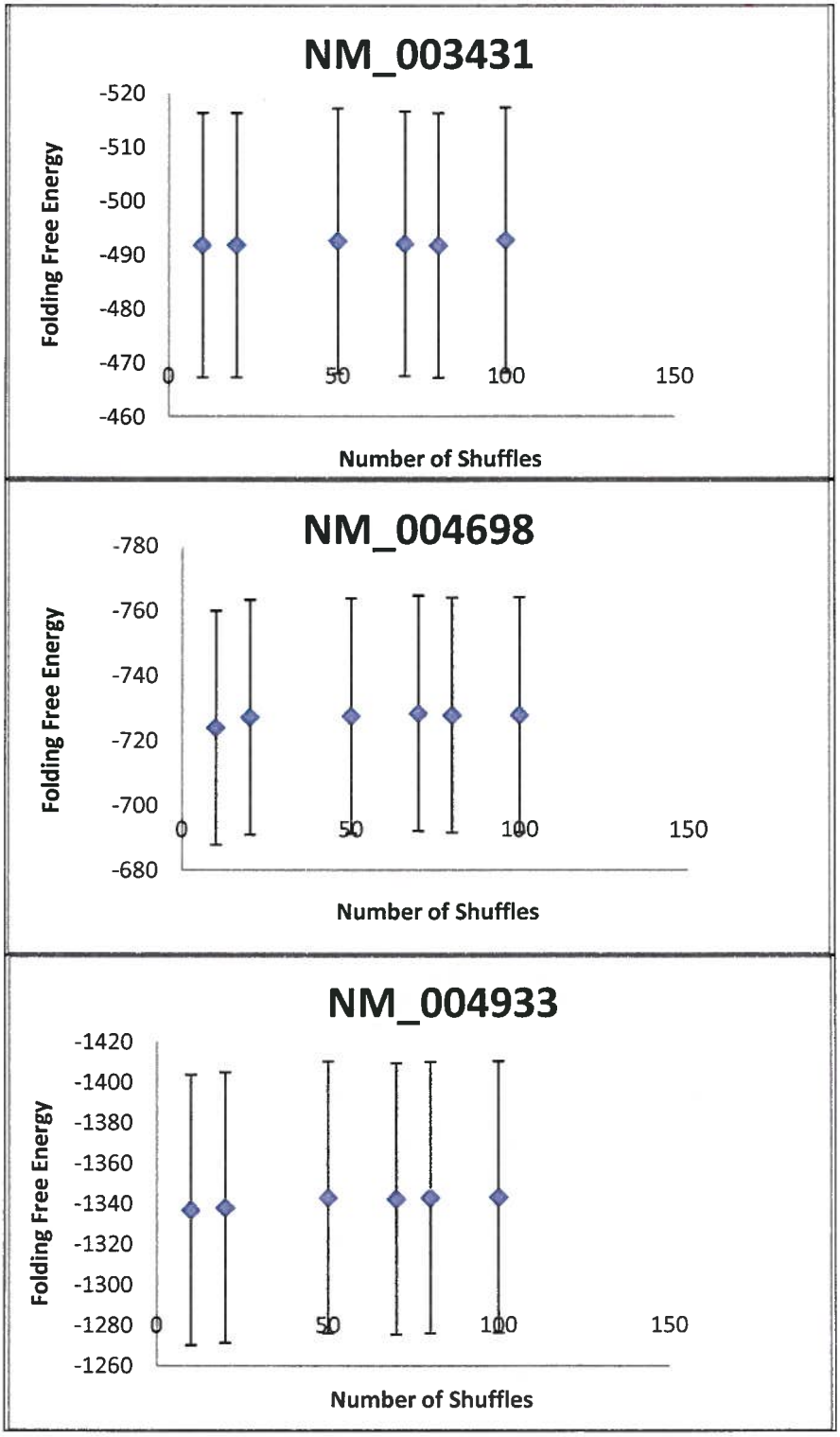

Figure 8. More genes displaying ten shuffles as the shuffle count needed to randomize a sequence.

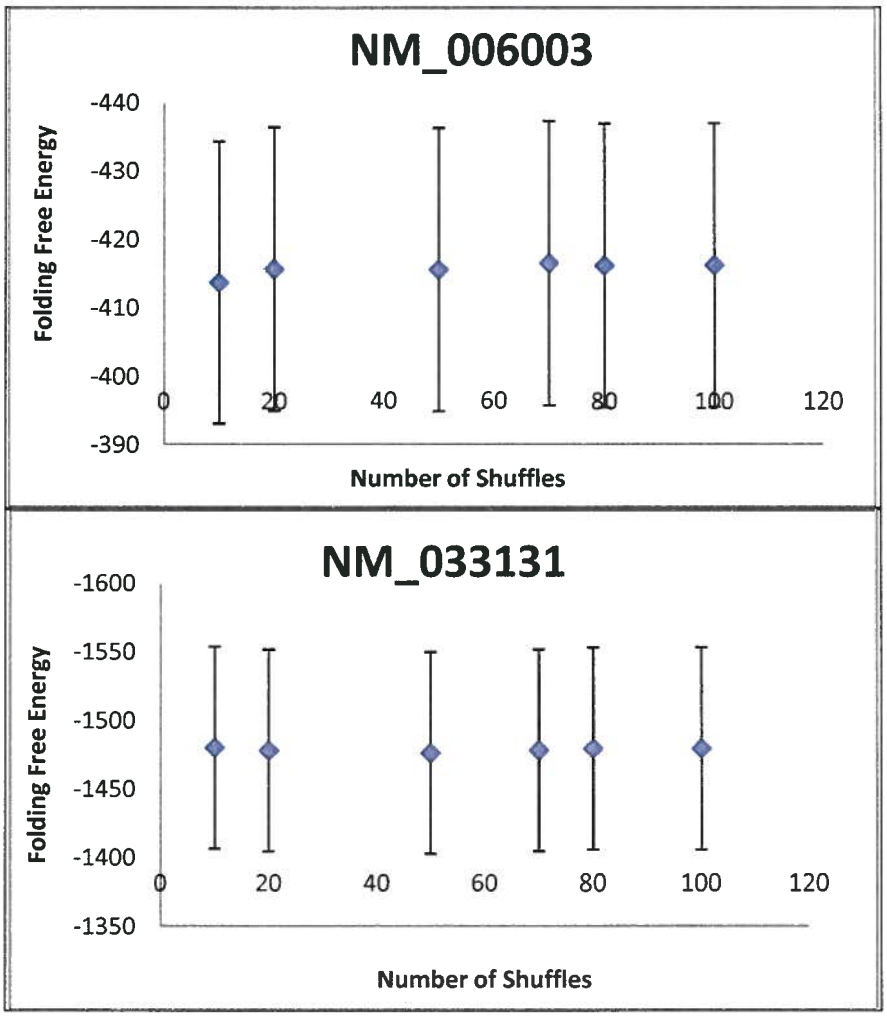

Figure 8. Ten (10) genes shuffled 100 times. With the addition of the error bars ten shuffles proves to be enough "shuffles" to randomize a sequence.

Folding Free Energies between different Version of RNAStructure

Before joining the Seffens Lab Group the version of RNAStructure used was 3.7. In the summer of 2006 the version of RNAStructure was updated to 4.2 versions to improve the computing performance and reduce time needed to fold sequences. This transition left us with different transcriptomes computed under different RNAStructure programs which made it difficult to compare the different species. Due to the change in

versions of RNAstructure used during this project, we assessed the impact this had on free energy calculations. A set of 100 human gene sequences were folded under both versions and the results plotted in figure 9. As seen from the regression equation, the free energies are offset by 33 kcal/mol between the two versions of RNAstructure. In addition to the human genes, we also folded sequences from plant, mouse and *strongylocentrotus p* transcriptomes to see if similar patterns exist. After several correlation experiments, the data proved that we could statistically convert energies of RNAStructure version 3.7 to version 4.2. The native energies and the randomized energies of the mouse transcriptome both versions gave a high correlation, 0.998 and 0.999 (figures 10 and 11) respectively; however, when comparing the Z-score of both versions the correlation was much lower (0.584) (figure 12). Since the Z-score is calculated as the (Native energy – Randomized energy)/ (Standard deviation), the only other variable to observe was the standard deviation. It was found that the standard deviation correlation between the two versions was 0.331 (figure 13) suggesting low correlation. Similarly to the human, other species behaved the same way with the native and randomized energies exhibiting strong correlations; while calculating the z-score as a whole number showed variation between the versions. Conversely, the standard deviation, native energy and randomized energy plotted independently of each other to possess convincing correlations between the two versions. The differences in the two versions could be attributed to the statistical equations used by the different versions.

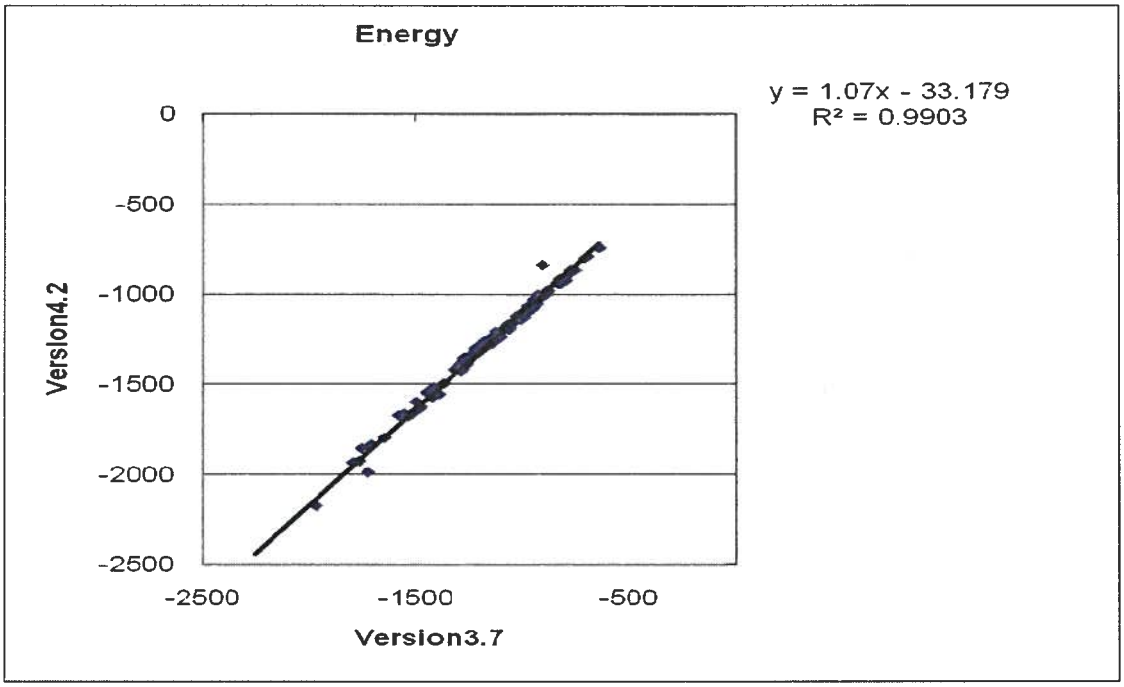

Figure 9. Difference in human sequences of folding free energy between version 3.7 and 4.2 of RNAstructure

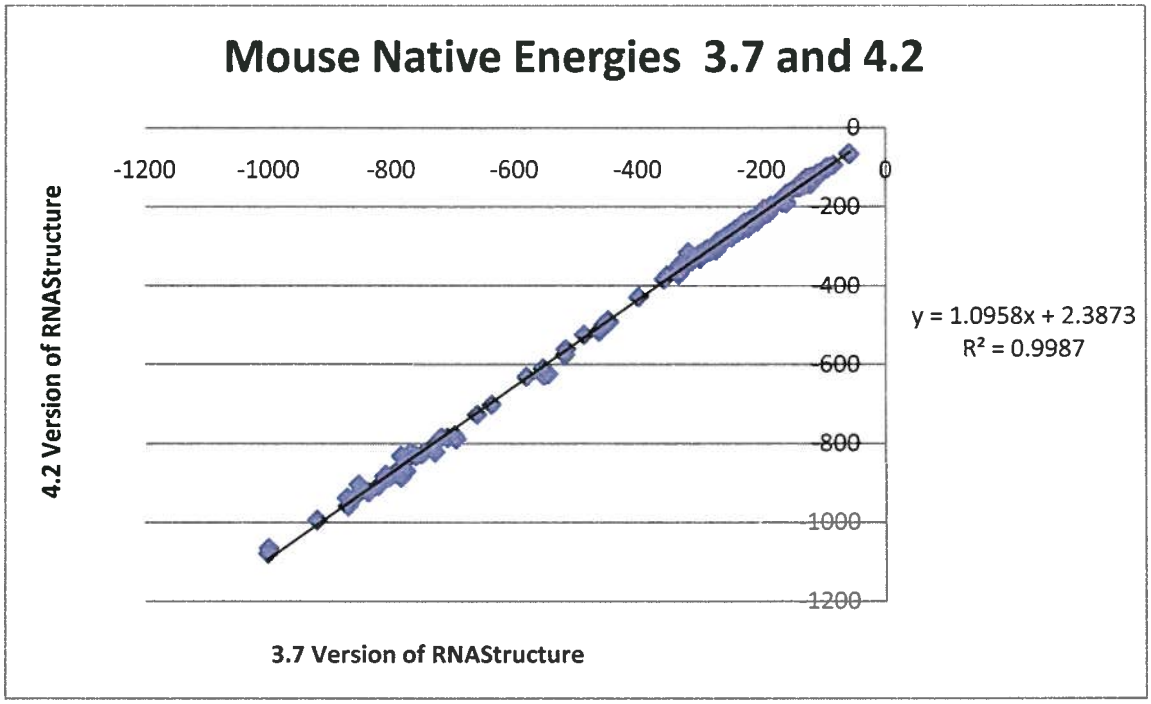

Figure 10. Native energies of mouse between 3.7 and 4.2 versions of RNAStructure. Close to a "perfect fit' between the native energies of both versions.

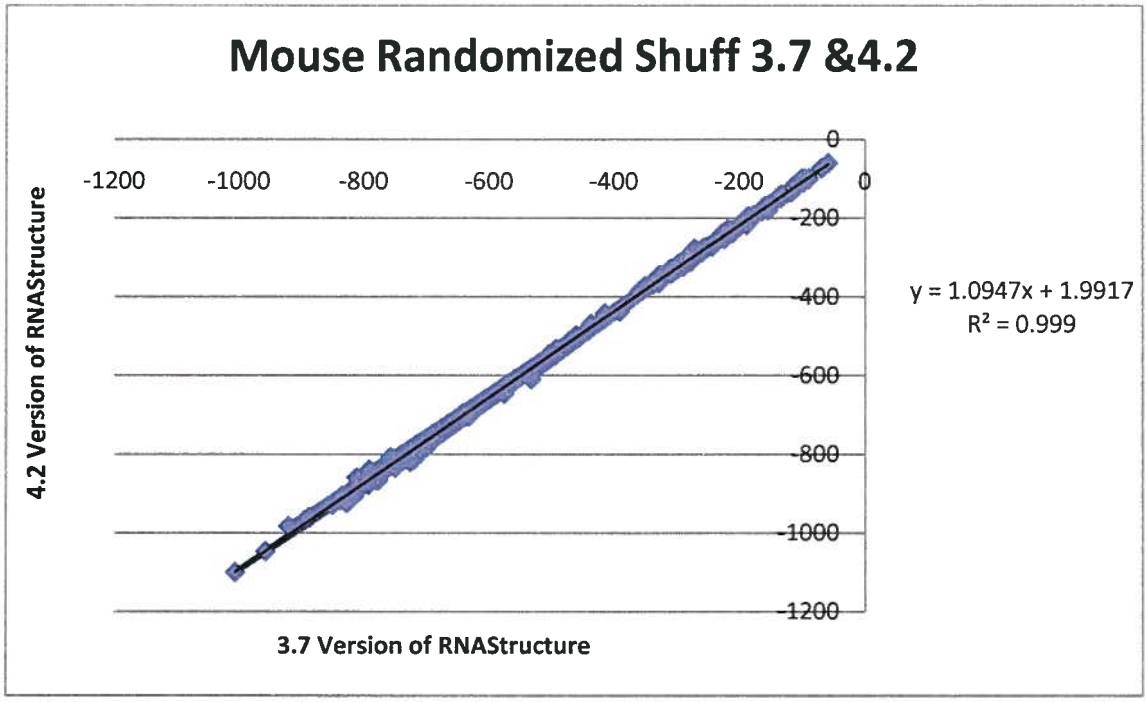

Figure 11. Randomized energies from mouse transcriptome displaying both 3.7 and 4.2 versions of RNAStructure.

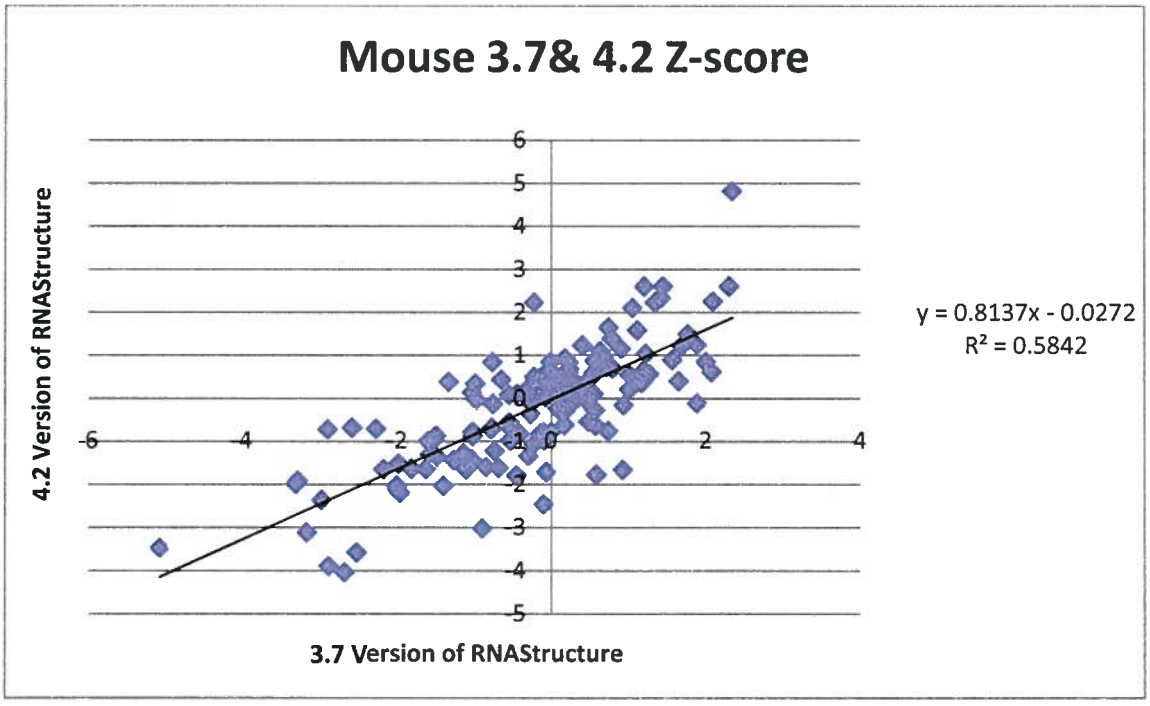

Figure 12. Plotted Z-score values from mouse transcriptome between both versions of RNAStructure.

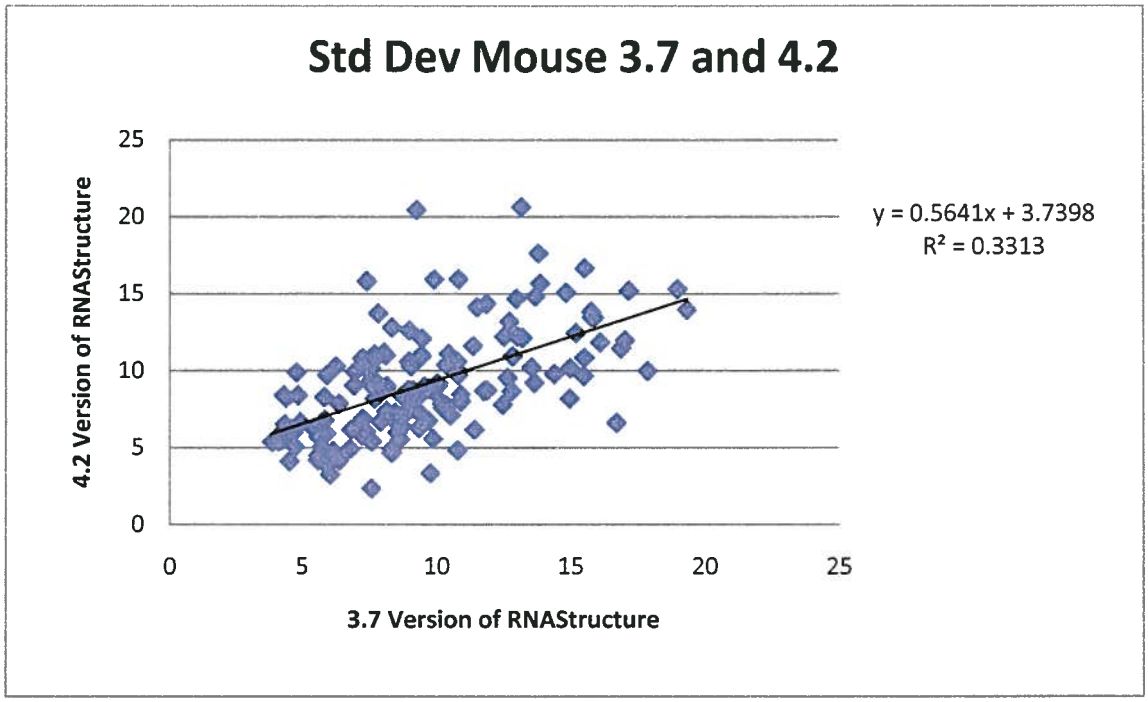

Figure 13. Calculations of standard deviations between both RNAStructure Versions

Gene Length and Folding Free Energy

Dr. Davis reported that mRNA folding free energy is dependent on sequence length. We investigated this finding by plotting calculated Z-scores as a function of sequence length. We performed a comparison study using the human, chicken, and chimpanzee transcriptomes to determine how important sequence length is to calculating the FFE's. It turns out that the Z-score is greatly affected by the sequence length because as the transcripts increase in length the FFE values become more negative and/or more thermodynamically favorable in terms of stability (figures 14 and 15).

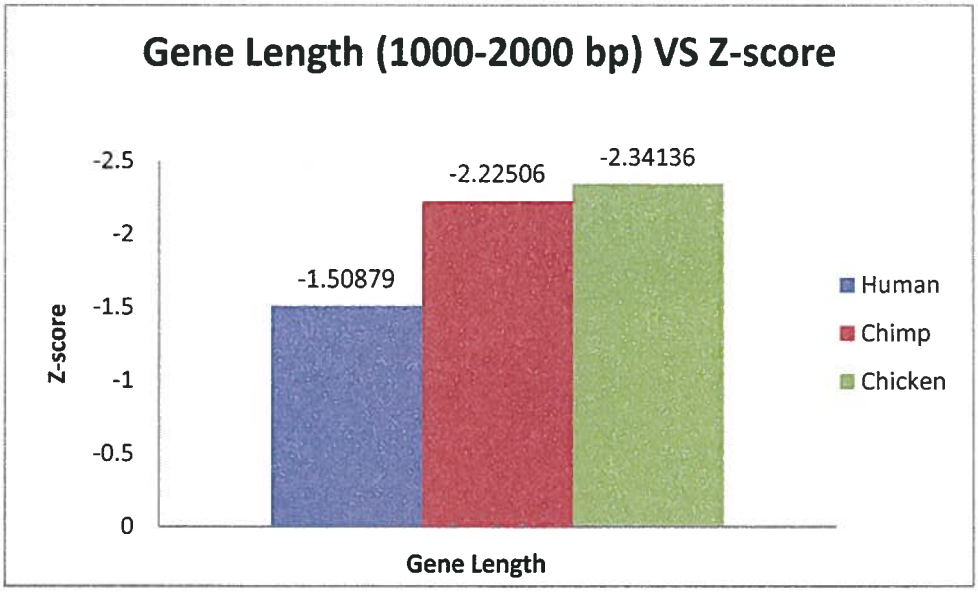

Figure 14. Z-score averages based on sequence length in different species.

It is widely accepted that natural selection favors shorter genic coding sequence length for higher transcriptional efficiency, for efficient protein synthesis, and for avoiding accumulation of deleterious mutation, on one hand, but evolution seems to improve the function of a protein through elongating its coding sequence on the other (Li 1997; Zhang 2000; Akashi 2003; Claverie and Ogata 2003; Wang, Hsieh, and Li 2005). Schneider and Ebert (2004) have recently argued that the covariation between genome size and gene length is expected to be strongest in smallest genomes and that selection for reduced gene length becomes progressively weaker when genomes become larger. Zhang (2000) observed that orthologous genes are longer in eukaryotes than in prokaryotes and that eukaryote specific proteins are longer on average than prokaryote-specific proteins. Xu et al indicated that the genic coding sequence has a relatively constant average length in both prokaryotes and eukaryotes in spite of the remarkable variation in the coding sequence length among individual genes within these genomes (Xu, Chen et al. 2006).

The coding sequence of a gene in the eukaryote kingdom is on average 445 bp longer than that in the prokaryotes. Xu et al suggests that natural selection has clearly set a strong limitation on gene elongation within the kingdom. We decided to normalize the transcriptomes based on sequence length to determine if sequence length affected the outcome of the calculated free energies computed (table 2).

Table 2. Normalization of transcriptomes based on sequence length.

Transcriptomes (Normalize based on sequence length)	Z-score	Z-score before normalization	% Neg. gene	Tot. sequences
Chicken	-2.34136	-2.45	93%	1965
Chimpanzee	-2.22506	-1.94	94%	197
Human	-1.50879	-1.66	79%	3059
Honeybee	-1.032	-0.895	74%	398
Strongylocentrotus	-1.012	-1.07	73%	411
Trypanosoma	-0.4712	-0.405	61%	429
Zebrafish	-0.3323	-0.3817	54%	1560
Mouse	-0.2687	-0.253	56%	2712
Plant	-0.2331	-0.197	55%	3947
T parva	-0.23059	-0.245	55%	237
Fly	-0.0276	-0.092	56%	2093

When normalizing the transcriptomes so that comparisons between organisms could take place we plotted the Z-scores of each species against the gene length of sequences within a range. Analysis of sequence length was done by creating bins with several ranges and each transcriptome was sectioned accordingly. The ranges of the bins were genes ranging from 1000 to 2000 base pairs, 2100 to 3000 bp, 3100 to 4000 bp, 4100 to 5000 bp, and 5100 and higher. Ideally we wanted each species to have at least 150 sequences for each bin. The range of choice was 1000 to 2000 bases per transcript and this was not difficult decide upon since some of the bins did not have adequate

number of sequences. For example, the bin that was normalized to sequence lengths that ranged from 2100 bp to 3000 bp had a few organisms that had large sample sizes for this range. Chimpanzee and *T. parva* both had a small number of sequences, 51 and 30 respectively and in the end we deemed this bin size inadequate. Other bins lacked a substantial number of species that did not have sequences with the length of choice. The *chicken* species was again the transcriptome that exhibited the most secondary structure as measured by Z-score displaying a value of -2.34136 (figure 14). However, the chicken did not contain the highest percentage of negative genes which was a direct correlation before we normalized the data (table 2). It would have been interesting to see what the percentage would have been if the chimpanzee had the same number of sequences as the chicken transcriptome.

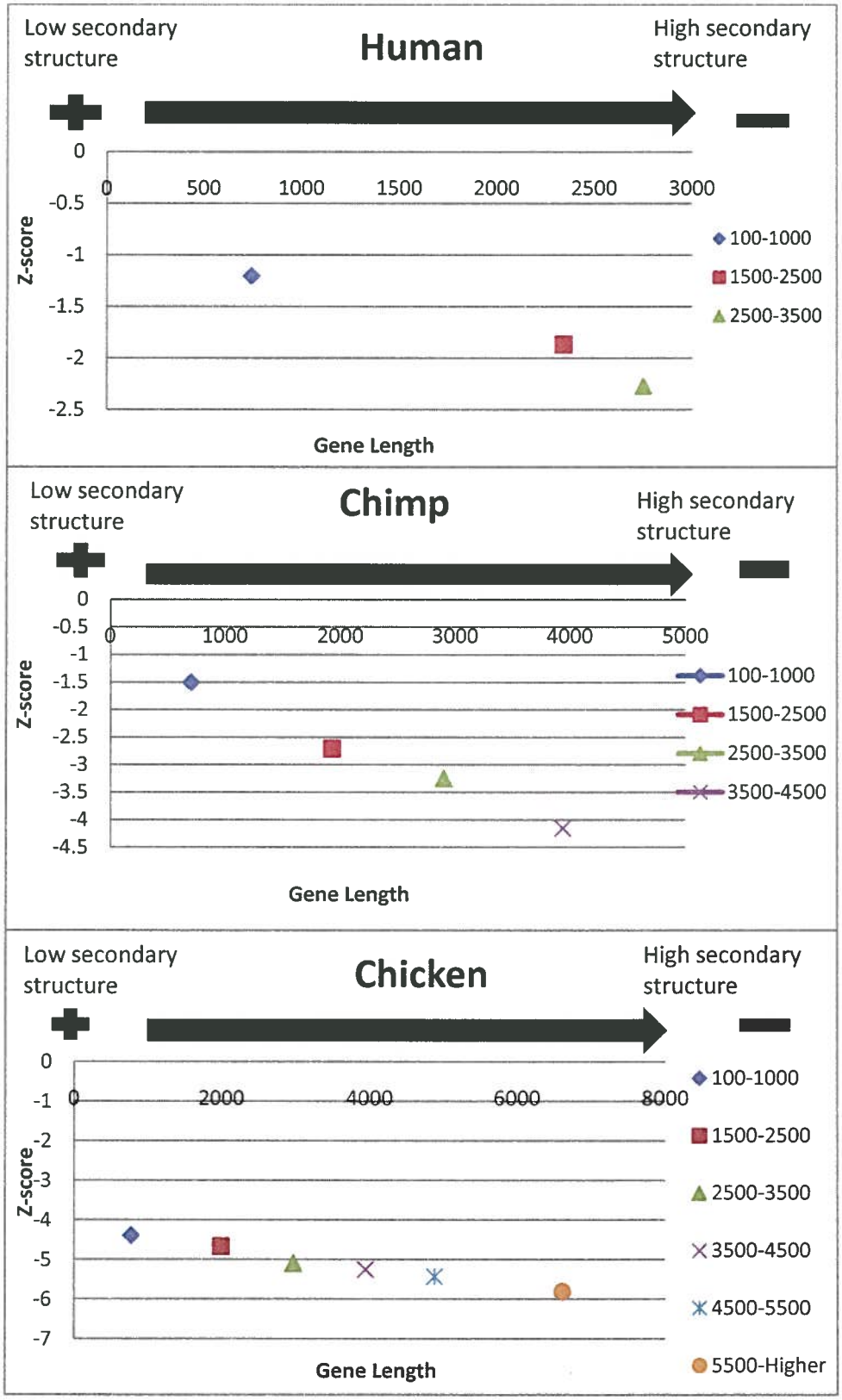

Figure 15. The human, chimpanzee and chicken transcriptomes as well as the others all exemplify increasing secondary structure as the length of the sequence increases.

CHAPTER 5

MUTI-REGRESSION ANALYSIS

Multiple regression is a flexible method of data analysis that may be appropriate whenever a quantitative variable (the dependent or criterion variable) is to be examined in relationship to any other factors (expressed as independent or predictor variables). Relationships may be nonlinear, independent variables may be quantitative or qualitative, and one can examine the effects of a single variable or multiple variables with or without the effects of other variables taken into account (Cohen, Cohen, West, & Aiken, 2003).

The question, why does certain species express more secondary structure as compared to others, is one that is complex and not easily solvable. Different species translate different genes which are specific to their habitat. Prokaryotes folding free energies generally show less secondary structure than eukaryotes because their base compositions are and their sequence lengths are usually shorter. Shorter sequence length limits the number of configurations a string of nucleotides can form. The hydrogen bonding exhibited in shorter sequences is not as prevalent as it is believe to be in larger sequences. Examining the relationship between dependent and independent variables such as dinucleotide base content and the percentage of negative genes representative of a transcriptome are factors that could possibly uncover some mysteries regarding excess secondary structure. For example, what is the correlation between the Z-scores and the percentage of negative genes comprised in a transcritptome?

We have already shown that gene length and secondary structure as characterized by Z-scores are directly proportional. (figures 14 and 15). Transcriptomes with large sequences on average contain more secondary structure than transcriptomes with shorter sequences. What is the correlation between the dinucleotide content and the Z-score?

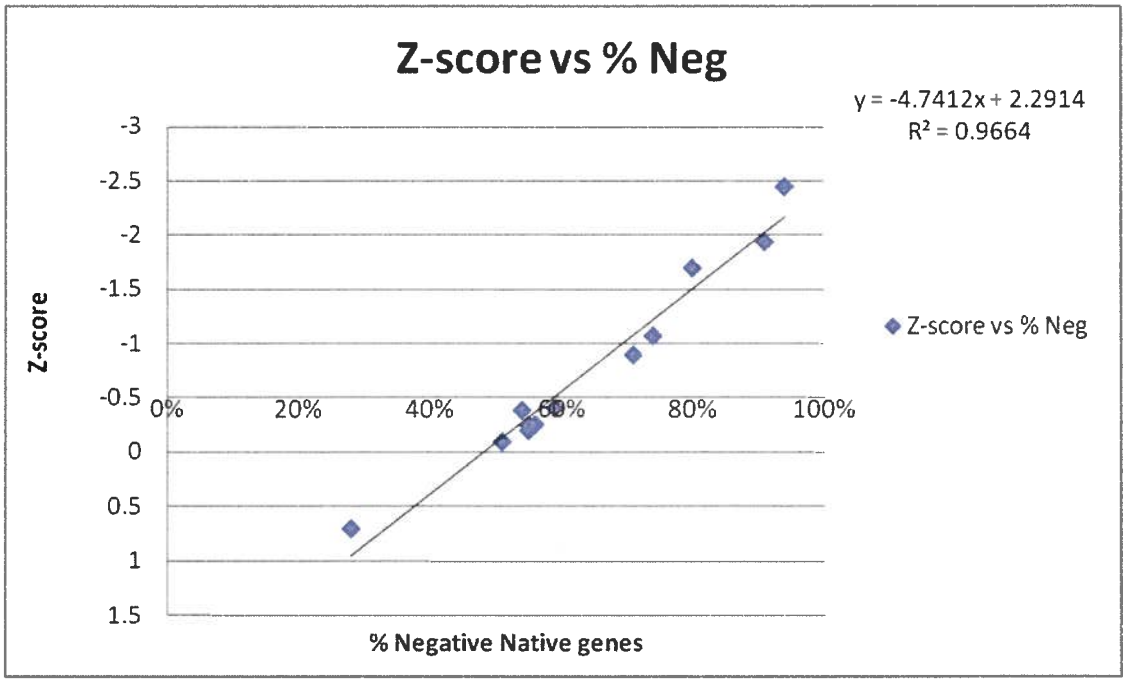

Figure 16. Z-score of all the organisms plotted against the percentage of negative genes of the transcriptomes

There's a direct correlation between the percentage of negative genes and the Z-score values of each of the organisms (figure 16). The amount of negative genes in a transcriptome is a key indicator of how much excess secondary structure is expressed. This is expected since the negativity represents thermodynamically favorable systems and so folding free energies with very negative values symbolize excess secondary structure. Naturally, transcriptomes with a large number of negative sequences will show a significant amount of secondary structure as depicted by figure 16.

Base composition is one of most fundamental features of a DNA sequence. It is given by the percentages of 4 different nucleotides, all taken on one strand. It is an observational fact that G pair with C and T pair with A generally holds, so we usually speak of G+C (or A+T) content (Fickett et al 1992). In an attempt to discover variables contributing to a more or less excess secondary structure we performed correlation studies between the Z-score and dinucleotide content of several organisms. There have been reports that dinucleotide composition has an influence on genomes. For instance, CpG islands (CGIs) are prominent in the mammalian genome owing to their GC-rich base composition and high density of CpG dinucleotides (Bird 1986). Most human gene promoters are embedded within CGIs that lack DNA methylation and coincide with sites of histone H3 lysine 4 trimethylation (H3K4me3), irrespective of transcriptional activity (Guenther 2007).

The chicken transcriptome used was comprised of 5,141 sequences and before the correlation was taken between the Z-score and dinulceotide content we first had to gather the correlation coefficients. The percentages of dinucleotide content and trinucleotide content both were calculated from RNAStructure folding program. Upon completion of folding, the Rgather program is executed which carefully lists all of the dinucleotide and trinucleotide percentages for that transcriptome. In Microsoft excel we sorted not only the dinucleotide pairing but its reverse compliment as well. For example, correlation coefficient was taken for the dinucleotide AA and its reverse compliment TT. Using Microsoft excel we were able to calculate the correlations between dinucleotide pairs and its reverse compliment. After correlations were taken for the sixteen pairings (AA, AG, AC, AT, CA, CC, CG, CT, GA, GT, GC, GG, TA, TG, TT, and TC) the correlation

coefficients were plotted against the Z-scores. The R^2, coefficient of determination was 0.0155 for the chicken transcriptome (figure 19). In regression analysis, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data. It provides a measure of how well future outcomes are likely to be predicted by the model. Figure 17 clearly demonstrate that the correlation between the Z-score and dinucleotide content is really weak. Also we did not find a dinucleotide frequency that was heavily prevalent in any of the species (table 3). We expected there to be a difference of dinucleotide frequency between species especially with the GC-content but that frequency was the most consistent between species.

Table 3. Dinucleotide frequencies calculated throughout the species. There were no outliers reported by the organisms.

Species	%AA	%CC	%CG	%GC	%GG	%TA	%TG
Human	7.63	7.75	3.23	6.83	7.31	3.64	7.74
Chimpanzee	7.74	7.36	2.94	6.58	7.11	3.68	7.83
Mouse	7.47	7.29	2.89	6.50	6.95	3.70	7.95
Chicken	8.21	5.97	2.95	6.60	6.45	4.12	8.13

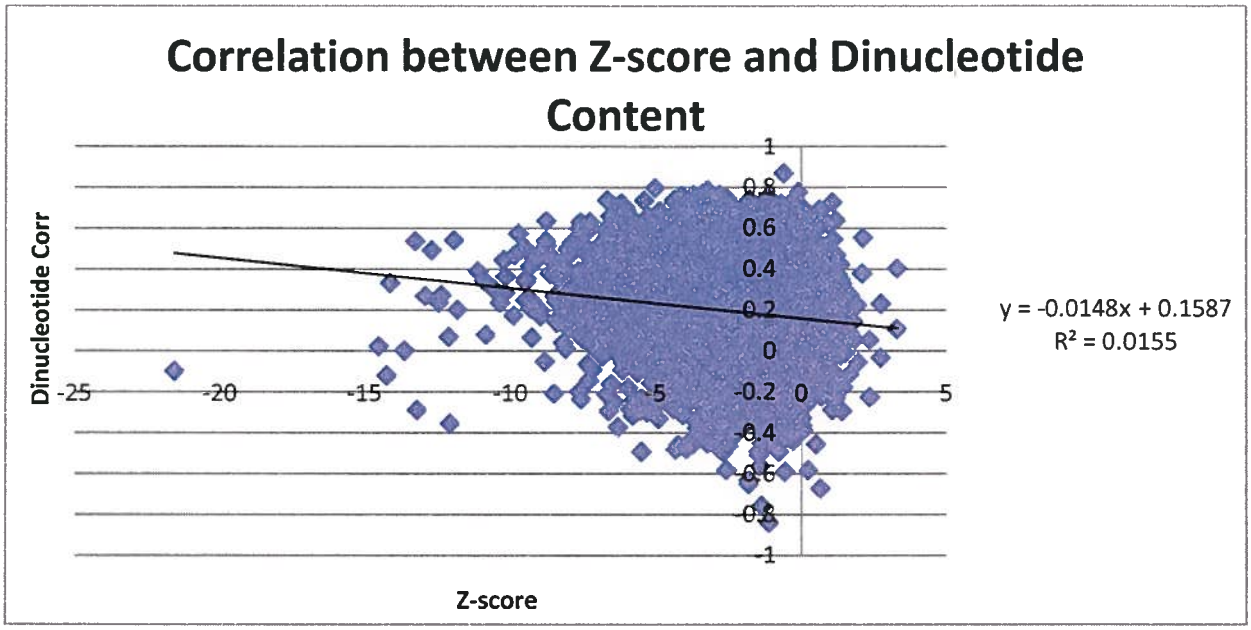

Figure 17. Chicken dinucleotide content plotted against the Z-score.

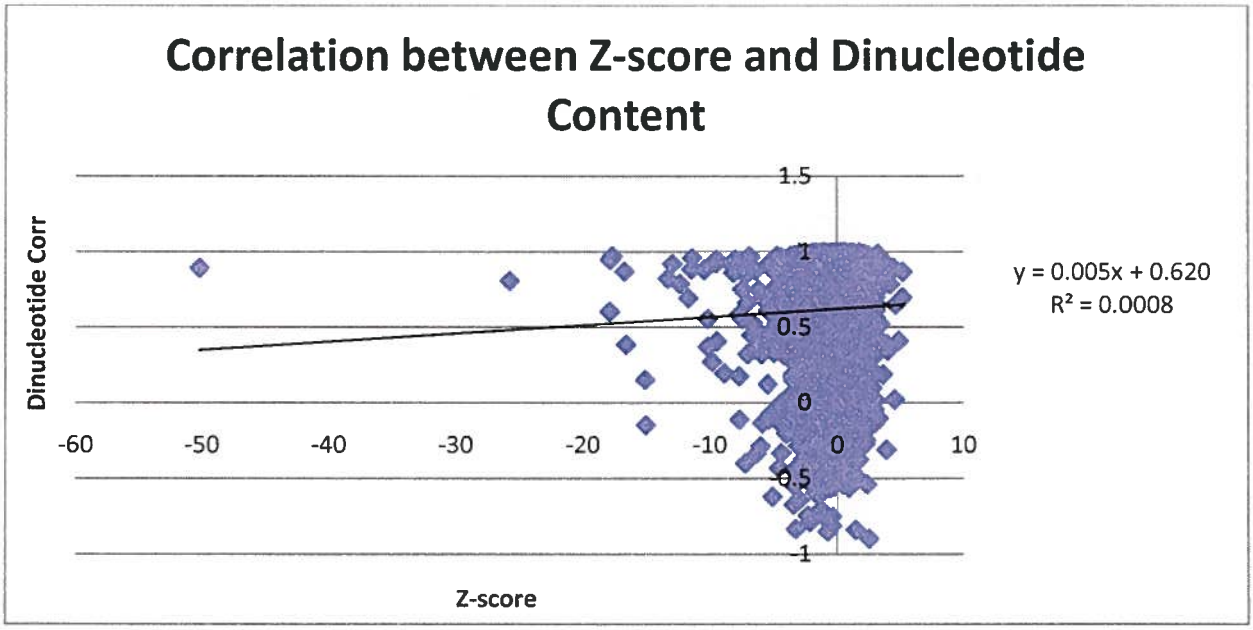

Figure 18. Human dinucleotide content plotted against the Z-score

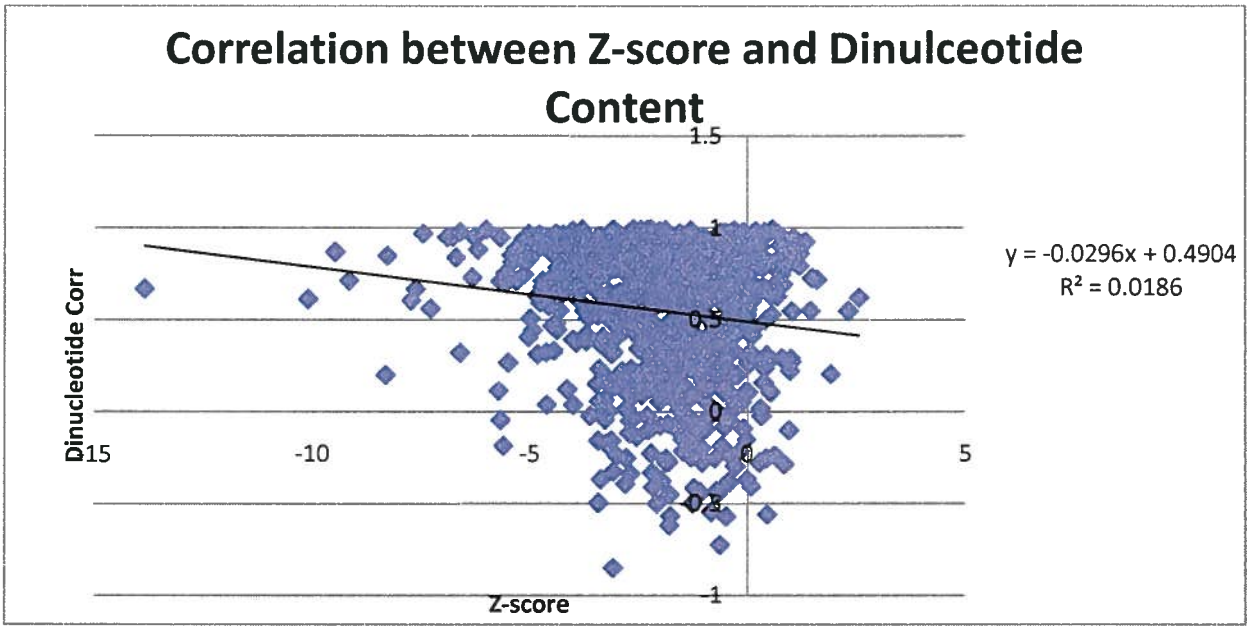

Figure 19. Chimpanzee dinucleotide content plotted against its Z-scores.

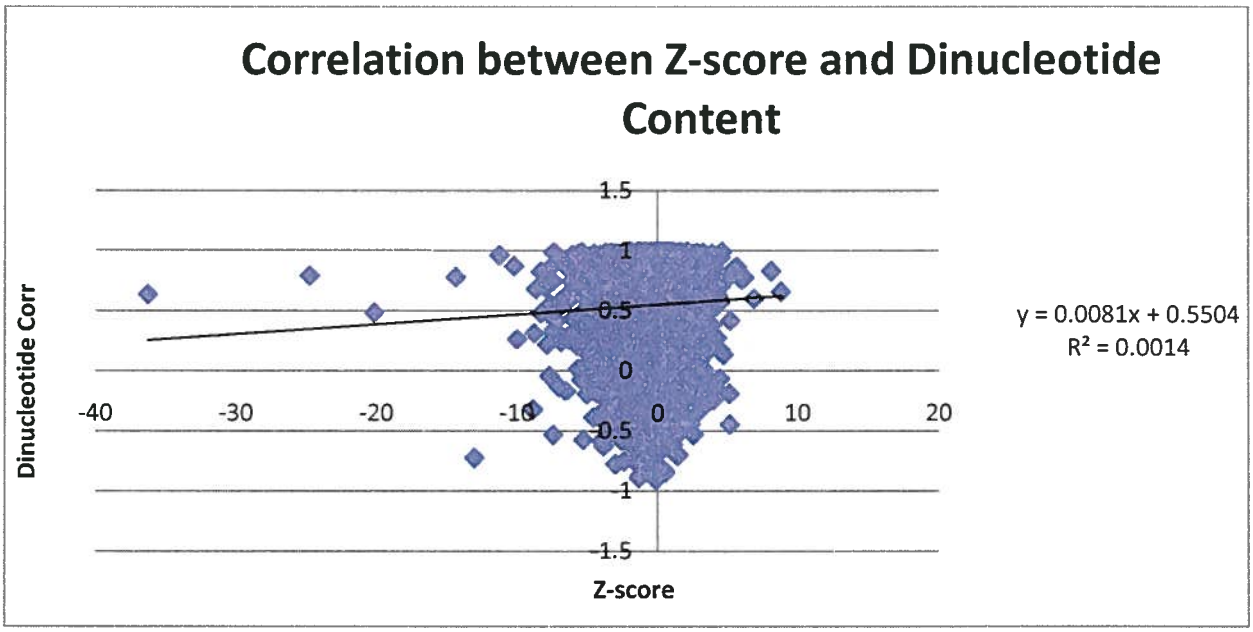

Figure 20. Mouse dinucleotide content plotted against the Z-score.

Again we notice the same trend for the human transcriptome that was comprised of 6221 sequences. The dinucleotide content does not seem to be an accurate predictor

for future outcomes. R^2 value computed for *homo sapien* was 0.0008 (figure 18). The chimpanzee transcriptome also reported a dismal R^2 value of 0.0186. Chimpanzee had total number of sequences folded for this analysis was 606 sequences (figure 19). The *Mus musculus* species had a total of 11,967 genes but again a really low R^2 value was reported (0.0014) (figure 20).

We also observed the trinucleotide content for the same species but the transcriptomes did not convey a good correlation. The human for example, gave a dreadfully low R^2 value of $3x10^{-5}$. Using the same gene sequences the trinucleotide content reported a lower R^2 values than the dinucleotide content values for all the species. The data for the trinucleotide content is not shown because of the dismal findings.

CHAPTER 6

NEURAL NETWOK ANALYSIS

Previously, a small neural network was trained on nucleic acid sequences to evaluate the network's ability to reliability and accurately predict codon usage (White G. and Seffens W., 1998). Various network architectures were used which depicted varying input units (nodes) that were synonymous to amino acids. This multi-layer (an input, an output and at least one hidden layer) perceptron (Figure 26) used back-propagation as its training algorithm and was successful in predicting 93% of the overall bases, 85% of the degenerate bases and 100% of the fixed bases in the best-trained network. The initial artificial neural network (ANN) simulator used for this work was called T-Learn ver1.0.1 (Plunkett K. and Elaman L., 1996). Pratt 2003 used this neural network (NN) simulator during the initial stages of this work. However, if the input was above a limit of 13 AA, T-Learn produced an error message referencing insufficient memory for our Binary 20 bit network encoding. Further investigation revealed that there was a limitation in the software that was not dependant on the installed hardware memory. Pratt then used a more robust simulator called Java NNS (Zell A., 1995) allowing AA window sizes of at least 20 (Figure 27). This increase in window size was important to encompass typical stem-loop structures that would be found in mRNA.

Peng produced evidence that there are long range correlations found within cDNA sequences when they are mapped onto a 'DNA-walk' (Peng *et. al.*, 1992). This evidence suggests that a region of a DNA sequence may have some degree of influence on a segment of DNA that is a considerable distance away from it. If mRNA sequences were optimized through evolution to yield greater negative free energies of folding, then there are more secondary structures than expected and consequently more stem-loop structures.

Neural Network Simulations

For the smaller NN training sets used by White and Seffens (1998), it was found that in order to arrive at or below an error level of $\leq 0.1\%$, the NN learning cycles ranged from 500 to 1200. However, the NNs used in this study were increased to a window size of 20 amino acids and had a data set of 200 sequences; which means that the networks and data set were more complex. Consequently, the number of training cycles or epochs increased substantially to between 200,000 to 250,000 cycles using backpropagation as the network-learning algorithm.

To optimize the performance of the network architectures surveyed, it was necessary to refine the various program parameters in order to produce networks that produced the best results for learning and predictive success. This was done by changing one parameter and holding the remaining parameters constant. These parameters included the number of units or nodes present in the hidden layer(s) of the network, the learning rate (μ), training set shuffle option and momentum (β). As the number of nodes in the hidden layer(s) were decreased the training error for the networks increased. Further when the number of units in the hidden layer(s) was increased beyond 100 units the networks learning performance started to degrade. This provides evidence to support the

conclusion that the optimum number of nodes was approximately 100 units, or five nodes per amino acid.

Additionally, Pratt 2003 examined the learning rate parameter for all the architectures. The learning rate (μ) refers to the rate of progression a NN moves towards a global minimum for learning. The program default value for μ was 0.2, and he adjusted this number >0.2 and <0.2 in increments of 0.05. Increasing the learning rate decreases the NN overall global error, and decreasing the learning rate had the opposite effect of increasing the overall NN global error.

Neural network learning parameters were assessed to determine the amount of influence on neural network performance. Sum square error logs recorded the learning error at specified generations. As neural network generations increased, overall the learning error decreased. Recent studies of the problem of training and generalization in neural networks have suggested (Denker 1987) that a critical number of examples exist, above which the generalization error falls off exponentially fast, due to a gap in the spectrum of generalization error (ε). In contrast, in the present work such a critical number does not exist. Instead, whenever a large number of genes can be learnt, ε approaches a power log. The power law behavior of ε is a manifestation of the absence of a gap in the spectrum of ε.

In an attempt to compare expected error and output error we established a control set that contained 1000 genes that was randomly selected and trained for one generation. After running the randomly selected training set for one generation we calculated the output error per amino acid. We then compared the output error with the expected error to

determine if the numbers were valid in terms of showing the neural network's capability of learning the data set. So for one generation the randomly selected data set computed a sum squared error of 6587.83 which gives 0.0475 errors per amino acid. Since there are four nucleotides each nucleotide base has a 1 in 4 chance or 0.25 probability of being correctly selected by the neural network. The expected initial error should be 0.25 x 3=0.75 per amino acid. This shows that from one generation the error drops from 0.75 to 0.0475 after one generation.

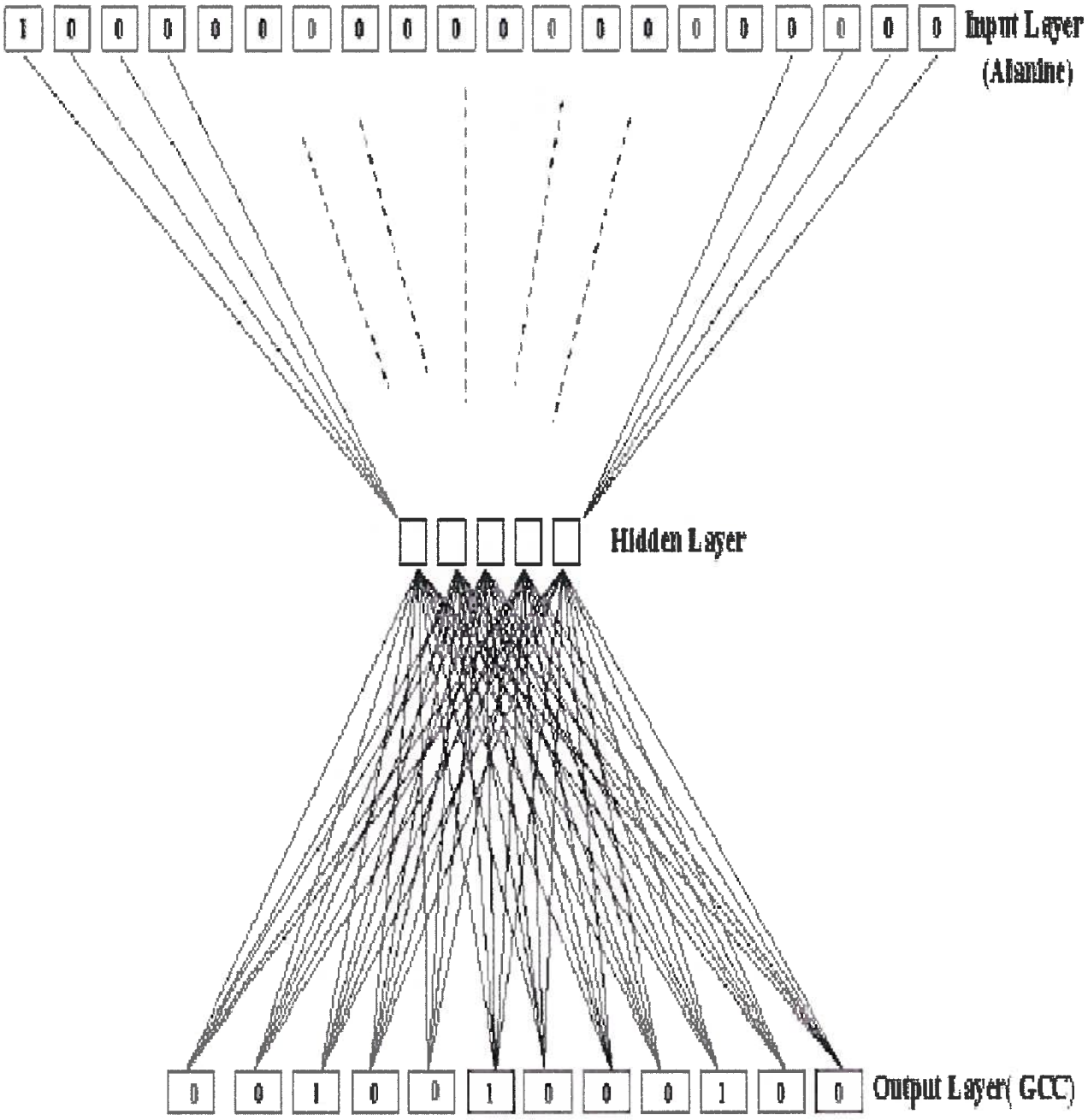

Figure 21. Binary 20-bit Encoding of NN Input Layer (Window size of 20 amino acids)

Initial training errors by Z-score sets

Mandy Lucas in 2006 used training sets that were examined during NN learning with runs of up to 10,000 generations (figure 22). NN learning parameters were

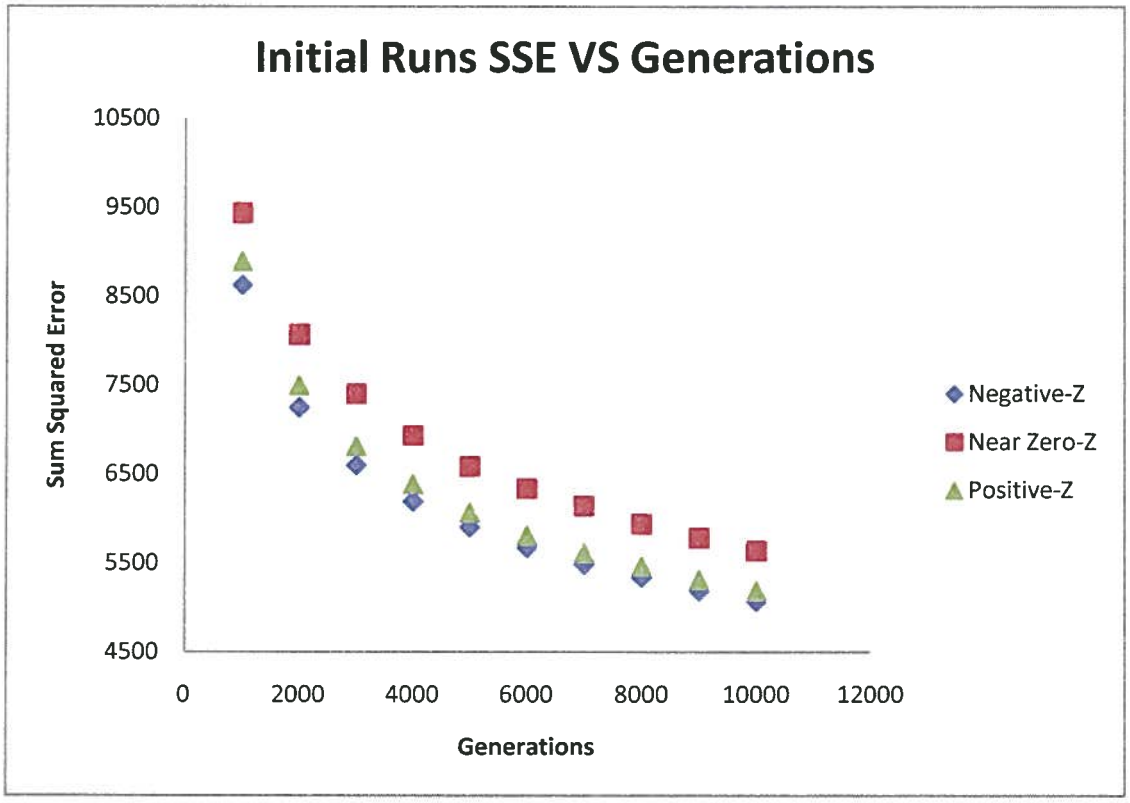

Figure 22. Sum Square Error Comparison. Initial runs show the Near Zero-Z training set to have the most error. The Negative-Z and Positive-Z training sets show similar progression. This is due to the presence of patterns in both sets and the lack theof in the Near Zero-Z training set. The error bars are too small to be seen above or below the marker.

assessed to determine the amount of influence on neural network performance. Sum square error logs recorded the learning error at specified generations. As NN generations increased, overall the learning error decreased. The Negative-Z training set tended to have more training error than the Positive-Z training set, while the training set with Near

Zero-Z had the most error (figure 22). Training runs were repeated three times out to 10,000 generations. The final errors were averaged and the standard deviation was calculated for statistical significance. The Negative-Z training set was not significantly different from the Positive-Z set runs.

After the preliminary studies Lucas decided to increase the number of generations to determine if similar results would occur again. Three subsets of 1000 genes were chosen from over 6000 folded human mRNA sequences but were pruned to eliminate redundant sequences. The subsets were placed into subsets according to Z-score, Negative Z, Near Zero Z, and Positive Z. The program "PattemMaker.exe" was placed into each directory with the isolated sequences. This program formats the sequences into training sets for the neural network program. Execution of "PattemMaker.exe" creates a pattern file, .pat. A neural network (JavaNNS) is trained using a large set of human mRNAs, in this case approximately 1,000.

The training sets were allowed to run up to 100,000 generations. Parameters were designated to specify the amount of neural network generations. The Error Log states the error at specified generations. As generations increased error decreased. Analysis of the Error Log for the Negative-Z, Near Zero-Z and Positive-Z subsets, the Near Zero-Z training set tended to have more error than the Negative-Z training set while the Near Zero-Z training set had the most error (figure 23).

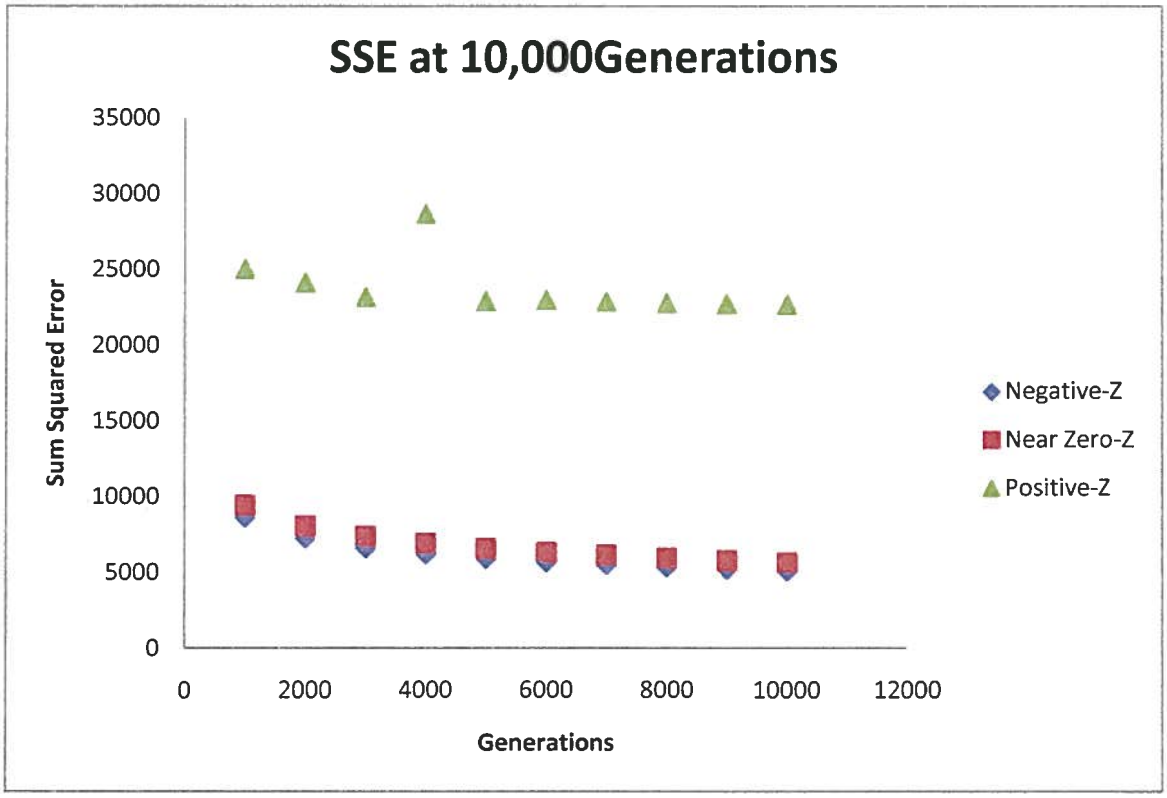

Figure 23. Sum Square Error Deviation. This depiction of sum squared error (SSE) at 10,000 generations shows the difference between the Negative-Z and Positive-Z training sets is not significant but the difference between the Near Zero-Z training set and the Negative-Z training set and the Positive-Z training set is significant.

Z-scores of human genes are a measurement of free folding energy. Human genes have a wide range of Z-scores. The training sets were obtained from a pool of genes with Z-scores from -52.82 to 5.12. The average was -0.48. The Negative-Z training set Z-scores were within the -52.82 to -1.79 ranges. They had a standard deviation of 2.75 and an average of-3.21. The Near Zero-Z training set Z-scores ranged from -0.58 to -0.07. The standard deviation for the set was 0.15 and the average was -0.32. Positive-Z training

set Z-scores range was 0.92 to 5.12. The set's standard deviation was 0.67 and its average was 1.65.

The chicken transcriptome exhibited very strong secondary structure throughout its genes and this is evident with the Z-scores computed. The Negative-Z set for the chicken had Z-scores ranging from -26.076 to -4.269 and illustrated an average Z-score of -6.0635 with a standard deviation of 1.934 within the subset. The Near Zero-Z set displayed Z-scores ranging from -2.744 to -1.846 with a standard deviation of 0.26252 and an average Z-score of -2.2908 within the subset. The Positive-Z set had Z-scores ranging from -1.003 to 3.74 and totaled an average Z-score of -0.1947 and a standard deviation of 0.70057 within the directory.

Z-scores of mouse genes ranged from -42.674 to 6.188 and followed the same trend as the other organisms as the low subset gave the most secondary structure and the high subset yielded the least amount of secondary structure expressed by the Z-score. The Negative-Z set had Z-scores ranging from -42.674 to -2.1437 and demonstrated an average Z-score of -3.267 with standard deviation of 1.937 within the subset. Near Zero-Z set gave Z-score values ranging from-1.0447 to -0.7273 with an average Z-score of -0.8827 and a standard deviation of 0.09739 within the directory. The Positive-Z training set expressed Z-score values ranging from 1.5302 to 6.188 with an average Z-score of 2.2353 and a standard deviation of 0.67746 within the subset.

The total set of Z-scores for zebrafish ranged from -200.197 to 6.199. The Negative-Z subset had Z-scores ranging from -200.197 to -1.631; while the Positive-Z subset contained Z-scores ranging from 0.768 to 6.199. The Negative-Z subset had an

average Z-score of -2.837 with a standard deviation of 5.491. Positive-Z subset has an average Z-score of 1.561 with a standard deviation of 0.7262. The Near Zero Z subset contained Z-scores ranging from -0.791 to 0.043. The subset had an average Z-score of -0.34954 with a standard deviation of 0.23703.

The human and mouse both displayed similar trends as the Negative-Z training set tended to have more training error than the Positive-Z training set, while the training set with near zero Z-scores had the most errors (Figure 24 & Figure 26 respectively). Chicken transcriptome follows the same trend as far the Near Zero-Z set displaying the most error recorded (figure 25). The Negative-Z training set was not significantly different from the Positive-Z set runs however; they were both significantly different from the Near Zero-Z training set.

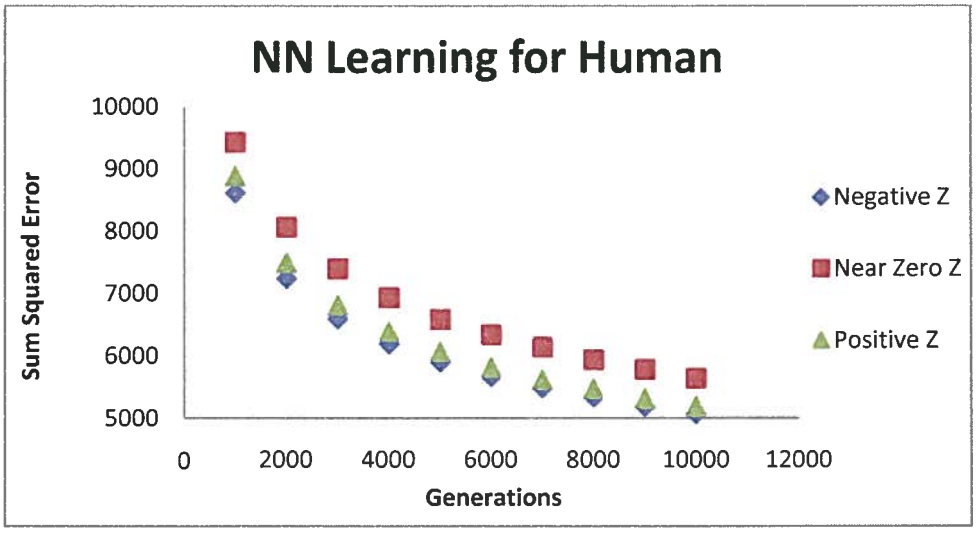

Figure 24. Training of neural network showing middle Z-score error to be higher than both the high and low Z-scores.

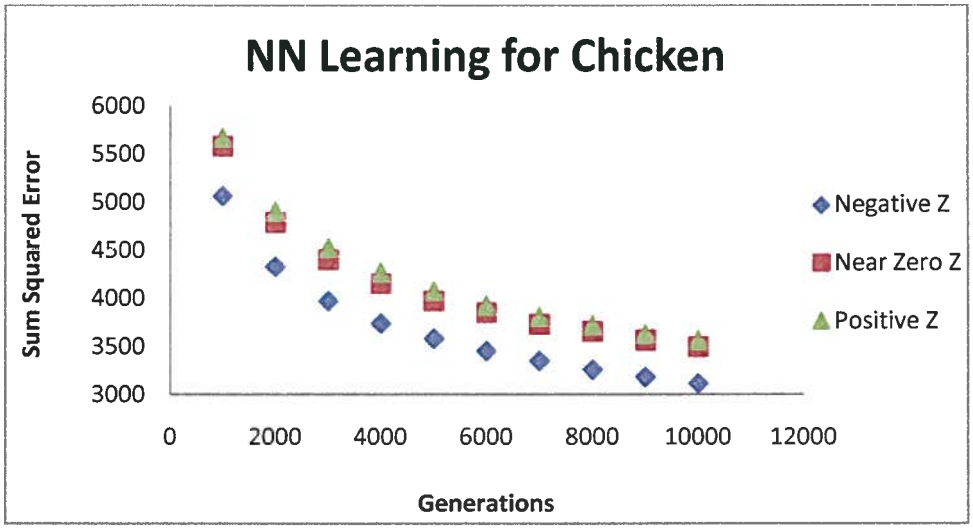

Figure 25. Learning of the chicken subsets. The negative-Z subset reported the least amount of errors.

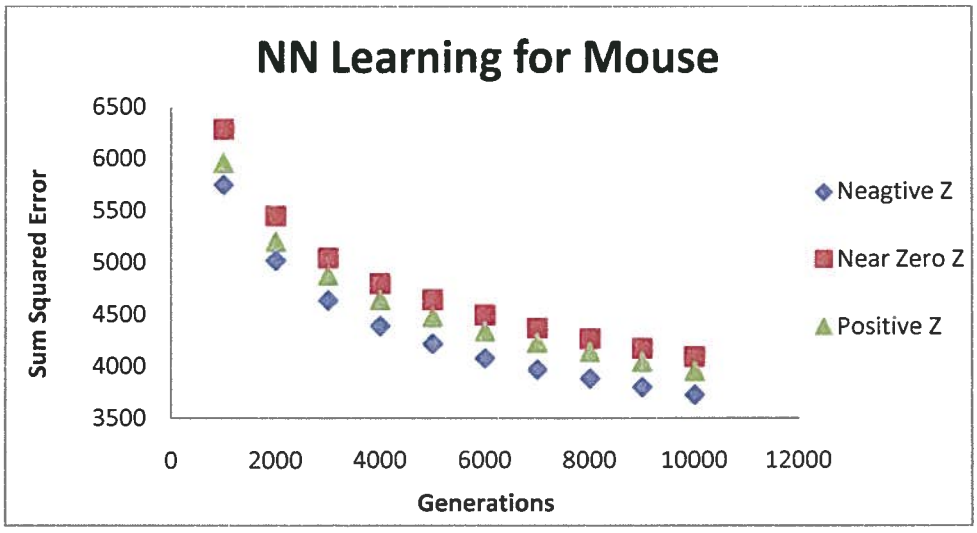

Figure 26. The positive-Z and negative-Z training sets show similar progression. This is due to the presence of patterns in both sets and the lack thereof in the middle training set.

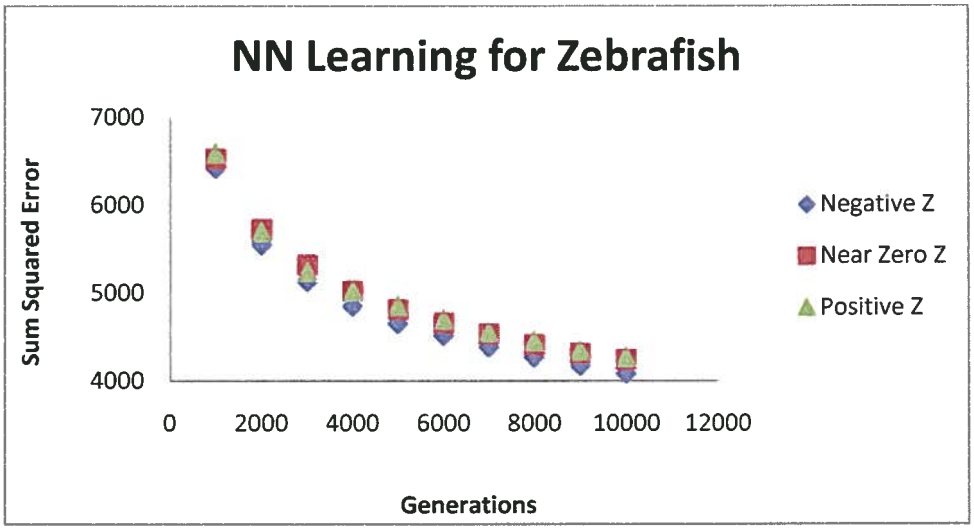

Figure 27. Zebrafish subsets displayed similar sum squared errors. However, the Negative Z subset did exhibit the lowest computed sum squared error.

All four species displayed the same trend where the directory with excess secondary structure displayed smallest number of sum squared error which means the NN 'learned' that subset the quickest (figure 28). The Near Zero-Z subset charted the most errors depicted by the SSE error log. What was intriguing was the fact that the Negative-Z and Positive-Z subsets (directory with excess secondary structure and directory little secondary structure) had very similar patterns as trained by the neural network. This report was not expected since both subsets contrast each other suggesting there are similar patterns in both the Negative Z' and Positive Z directory.

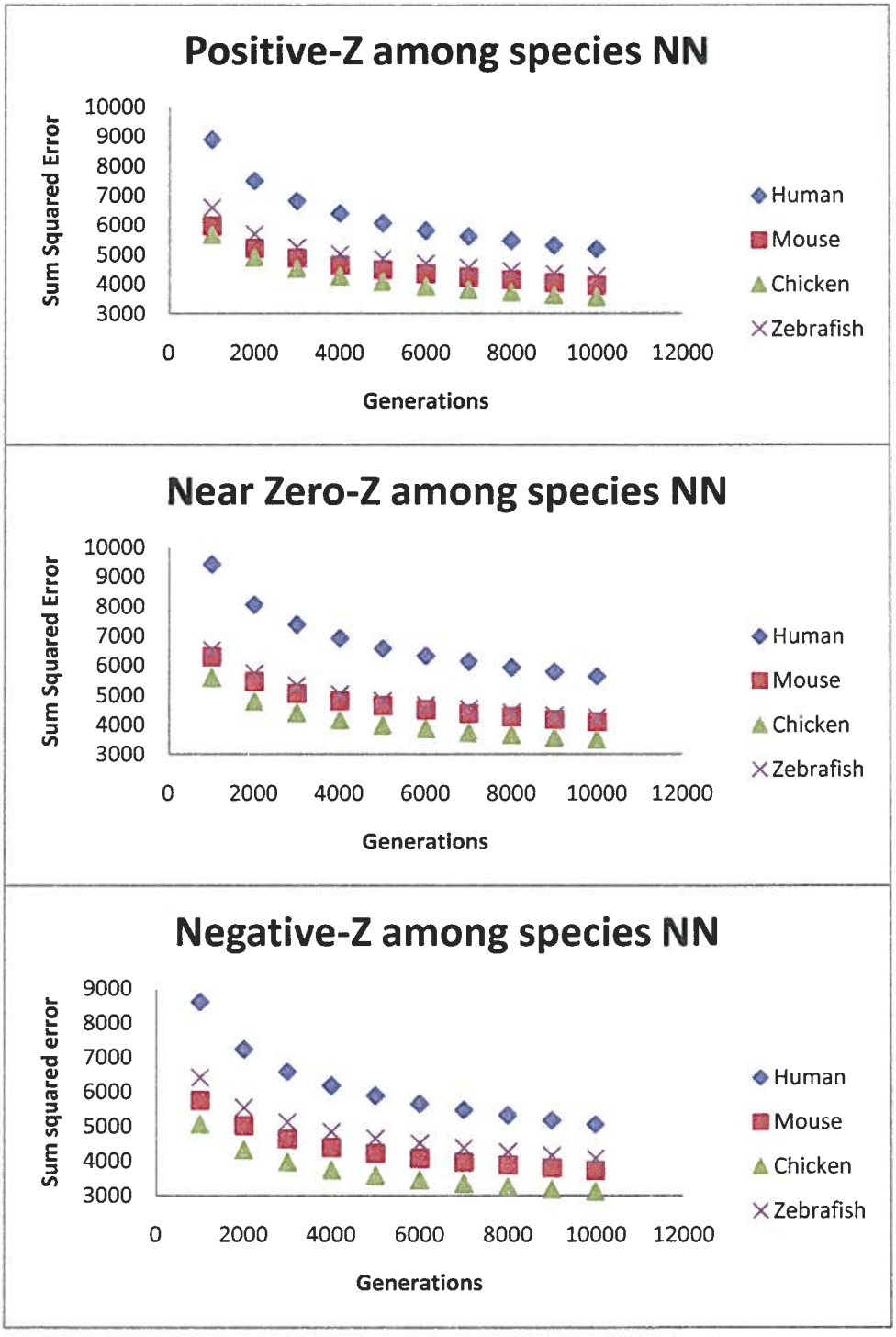

Figure 28. The chicken transcriptome exhibited the lowest amount of sum squared errors for each of the directories. The larger sequences are believed to contribute to this trend. Sequence length is proportional to the amount of secondary structure contained in the transcriptome.

CHAPTER 7

DISCUSSION

Comparative Study

Genes can be classified according to whether they are more or less stable in calculated folding free energy compared to randomized sequences. Examination of the global mRNA secondary structure of transcriptomes reveals the presence of excess free energy in native or wild-type sequences compared to randomized sequences (Digby and Seffens, 1999) due to more mRNA secondary structure typically in the form of stem-loops (Tables 1 and 2). Since it is proven that structural RNA features are caused by complementary base-pairings it is believed that secondary structure is involved in the regulation of mRNA degradation (Jacobson *et. al.* 1998). In addition, excess RNA secondary structure displayed in native sequences may be involved in gene regulation mechanisms, intron splicing, or steady state mRNA levels.

Dr. Davis concluded that the human transcriptome is more stable than the randomized mRNAs which suggested that evolution has produced gene sequences that transcribe into messages that contain more secondary structure than expected. However, *Homo sapiens* were not the only species where the native sequences had more excess free energy than the randomized set. Examination of the mRNA secondary structure of *Mus musculus* (mouse), *Gallus gallus* (chicken), *Pan troglodytes* (chimpanzee),

Stronglocentrotus purpuratus, Trypanosoma brucei, *Danio rerio* (zebrafish), *Drosophilia melanogaster* (fruit fly), *Apis mellifera (honeybee), Theileria parva* (protozoa) and *Arabidopsis* transcriptomes reveals the presence of excess free energy compared to randomized sequences. Most of the mRNA structure contributing to global free energy of folding is located within the coding region of the transcripts due to the smaller size of the UTRs. Hence the coding sequence in genes may result from a multi-objective optimization process in evolution involving both protein and mRNA sequence structures.

But when analyzing the transcripts of the chicken species we observed that this transcriptome was comprised of very large sequences. In fact, when observing the sequence length of the transcriptomes to normalize the data, the chicken transcriptome had sequences that contained more than six thousand bases. One suggestion for the large sequences analyzed could be due to the version of RNAStructure used to fold the sequences. As reported earlier the later version of RNAStructure (4.2) generated FFE's that were more negative as to compare to the older 3.7 version. Another possible reason why the chicken transcriptome displayed the lowest amount of sum squared error could be due to a lack of complexity in the sequence patterns. The chicken sequence patterns detected by the neural network were easier to train because of the arrangement of bases positioned in the sequence were readily learnable. The opposite can be said for the human subsets when analyzing the neural network progression. We believe the human transcriptome contained more complex sequence patterns which caused the neural network to generate more errors as compared to the other species. It would be interesting to see training sets comprised of orthologs across species to analyze the training of the

genes. This experiment would give insight as to why certain species generate certain FFE's.

This research has allowed us to determine a shuffle count to randomize an mRNA sequence. The controversy surrounding the number of shuffles needed to shuffle a gene was solved by folding ten different genes after they were shuffled 100 times. As mentioned previously, only two genes out of ten required 50 shuffles for calculation of a reliable Z-score. However, inspection of the remaining eight genes suggests that ten shuffles were sufficient to calculate Z-scores. With this analysis we felt confident when selecting genes based on their Z-score value for input into the neural network. Knowing that the sample size and the correct number of shuffles were statistically correct this added validation to our approach for analyzing sequence patterns in mRNA sequences.

Correlation studies between the dinucleotide and trinucleotide content and secondary structure characterized by Z-scores computed very low correlations. The statistical relationship between dinucleotide or trinucleotide content and the Z-score reported dismal results. However, this does not prove that base content does not have influence with folding free energies but it does tell us that the Z-score may not be the preferred statistical method for characterizing transcriptomes. But it is interesting that two variables (Z-score and base content) that are both associated with folding free energy did not relate at all. We did analyze dinucleotide frequencies between the species but there was no dinucleotide frequency outlier which poses questions for reasoning (Table3).

Neural Network

As mentioned previously, different species display different FFE's and as a result have different amounts secondary structure. From this evidence the decision to investigate the patterns in mRNA sequences to find out if there is tendency or bias between species characterize with excess secondary structure and transcriptomes that show the least amount of secondary structure. The subset with negative Z-scores showed the lowest neural network sum squared errors. This indicates that there are more detectable sequence patterns in genes with more secondary structure than in genes with Positive Z-scores. It appears that Negative and Positive Z-score sequences have patterns in codon usage that are more favorably detectable by neural network training (Figure 23). For example, the training sets in both the human and mouse seem to show very similar learning progression (Figures 26 and 28). One reason for this result could be Negative-Z and Positive-Z sequences both have specific patterns that give rise to low or high amounts of secondary structure. Near Zero-Z values had a higher initial error rate than the other two extremes presumably because an average gene has less specific patterns than the other two extremes. These results furnish evidence which support our hypothesis that the folding free energies of mRNA's within transcriptomes does affect the sequence patterns in genes when sorted based on secondary structure. Not only are there differences between the amount of secondary structure exhibited but there were differences in the error between species as well. The fact that the NN generated differences in sum squared error between species is not all that surprising but why does human sequences have higher SSE's as oppose to the species for each of the three subsets (Figure 28)? The neural network learning and predictive capabilities is due to

arrangement of patterns in gene sequences. Even though the chicken transcriptome displayed the most secondary structure characterized by Z-score values, we propose that the human have a more complex arrangement of bases which causes the NN to generate more errors per generations.

CHAPTER 8

CONCLUSION

Finally, we conclude that, across the sampled subsets of the transcriptomes there is a significant difference in the overall folding free energies of the native sequences and the randomized sequences. Therefore, the transcriptomes on average are more stable than the randomized sequences. In terms of secondary structure and complexity, there does seem to be a correlation between the two variables. However, the chicken did display more secondary structure by displaying more negative genes than positive genes in its transcriptome than the chimpanzee and the human. One reason the native sequences display a more negative energy could be nature's selection for greater transcript stability in the majority of human genes. It seems that the native sequences have evolved in such a way as to form stable mRNAs to protect them from degradation.

Structural RNA features created by base pairing have been implicated in the regulation of mRNA degradation (Jacobsen *et al* 1998). Such structures are believed to be a part of the overall global free energy of folding for an entire mRNA sequence. Seffens and Digby stated that the global free energy of folding arises from the coding sequence and selection of codons. Investigation of free energy displayed in mRNA sequences has been studied. Comparison between native or original sequences with randomized or re-shuffled sequences has been determined. Analysis of folding free energy of mRNA sequences as a result of specific sequence patterns has not yet been determined. Are the

folding free energies present in specific mRNA sequences the product of specialized patterns noticed in genes? For example, does the chicken transcriptome have more secondary structure because it contains more sequence patterns. In observing this comparative study we cross linked several transcriptomes and their corresponding Z-scores with the pattern recognition capabilities of an artificial neural network. The data produced in this experiment provided evidence which support our hypothesis that the folding free energies of mRNA's within transcriptomes does affect the sequence patterns in genes when sorted based on secondary structure. There are differences between the amount of secondary structure exhibited and there were differences in the error between species. Sum squared error measurement of neural network training allows the magnitude of patterns in codon usage to be determined. Error differences showed more patterns led to better prediction. Errors in prediction were relative to the overall presence of the amino acid in the validation set of genes. With the assistance of a folding program and neural network we were able to show that patterns exist in mRNA that may be responsible for excess secondary structure.

The chicken transcriptome displayed the most secondary structure characterized by Z-score values but generated the least amount of sum squared errors as computed by the NN of all the subsets. The human transcriptome contained large sequences just as the chicken transciptome but generated the most errors. The zebrafish and the mouse transciptomes both were trained by the neural network and produced a similar learning progression curve. We propose that the human have a more complex arrangement of bases which causes the NN to generate more errors per generations. Base content and composition maybe more intricately place within a sequence of a *homo sapien*. The

chicken transcriptome had larger sequences than the human but we believe the reason the neural network computed less errors is due base arrangement and in some cases base orientation. The chicken transcriptome lack complexity in their mRNA sequences and as a result allowed the NN to quickly learn the training set. We believe the same goes for the zebrafish and the mouse transciptome as well that because of the lack of complexity displayed in their mRNA sequences the neural network generated a favorable learning progression.

BIBLIOGRAPHY

Akashi, H. (2003). "Translational selection and yeast proteome evolution." Genetics 164(4): 1291-303.

Auweter, S. D., F. C. Oberstrass, et al. (2006). "Sequence-specific binding of single-stranded RNA: is there a code for recognition?" Nucleic Acids Res 34(17): 4943-59.

Arms, K and Camp, P. (1987). Biology. "DNA and Genetic Information."3rd Ed. Saunder College Publishing. Orlando, Florida. 168-193.

Ball, L. A. (1973). "Mutual influence of the secondary structure and information content of a messenger RNA." J Theor Biol 41(2): 243-7.

Ball, L. A. (1973). "Secondary structure and coding potential of the coat protein gene of bacteriophage MS2." Nat New Biol 242(115): 44-5.

Berg, J, Tymoczko, J and Stryer, L. (2002). Biochemistry. "RNA synthesis and Splicing." 5th Ed. W.H. Freeman. New York, New York.

Berg, J, Tymoczko, J and Stryer, L. (2006). Biochemistry. "RNA synthesis and Splicing." 6th Ed. W.H. Freeman. San Francisco, California.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell 125(2): 315-26.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature 321(6067): 209-13.

Bloomfield, V.A., Crothers, D.M., Tinoco, I. (1974). "Physical Chemistry of Nucleic acids." Harper, New York.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R, Lautrup, B., Norskov, L., Olsen, O. And S. Petersen(1988). Protein Secondary Structure and Homology by Neural

Networks: The α-helices in Rhodopsin. FEBS Letters Vol. 241 No. 1,2 pp. 223-228.

Buckanovich, R. J. and R. B. Darnell (1997). "The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo." Mol Cell Biol 17(6): 3194-201.

Clausen-Schaumann, H., M. Rief, et al. (2000). "Mechanical stability of single DNA molecules." Biophys J 78(4): 1997-2007.

Claverie, J. M. and H. Ogata (2003). "The insertion of palindromic repeats in the evolution of proteins." Trends Biochem Sci 28(2): 75-80.

Crick, F. H. (1954). "The Complementary Structure of DNA." Proc Natl Acad Sci U S A 40(8): 756-8.

Delbruck, M. (1954). "On the Replication of Desoxyribonucleic Acid (DNA)." Proc Natl Acad Sci U S A 40(9): 783-8.

Delisi, C. and D. M. Crothers (1971). "Prediction of RNA secondary structure." Proc Natl Acad Sci U S A 68(11): 2682-5.

Demuth, J. P., T. De Bie, et al. (2006). "The evolution of mammalian gene families." PLoS One 1: e85.

Dolinnaya, N. G. and J. R. Fresco (1992). "Single-stranded nucleic acid helical secondary structure stabilized by ionic bonds: d(A(+)-G)10." Proc Natl Acad Sci U S A 89(19): 9242-6.

Dong, F., H. T. Allawi, et al. (2001). "Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA." Nucleic Acids Res 29(15): 3248-57.

Dotu, I., W. A. Lorenz, et al. (2009). "Computing folding pathways between RNA secondary structures." Nucleic Acids Res.

Doty, P., H. Boedtker, et al. (1959). "Secondary Structure in Ribonucleic Acids." Proc Natl Acad Sci U S A 45(4): 482-99.

Fitch, W. M. (1974). "The large extent of putative secondary nucleic acid structure in

random nucleotide sequences or amino acid derived messenger-RNA." J Mol Evol 3(4): 279-91.

Fitch, W. M. (1983). "Calculating the expected frequencies of potential secondary structure in nucleic acids as a function of stem length, loop size, base composition and nearest-neighbor frequencies." Nucleic Acids Res 11(13): 4655-63.

Fong, P. (1967). "The transcription of the DNA molecule." Proc Natl Acad Sci U S A 58(2): 501-5.

Fournier, M. J. and H. Ozeki (1985). "Structure and organization of the transfer ribonucleic acid genes of Escherichia coli K-12." Microbiol Rev 49(4): 379-97.

Freyhult, E., P. P. Gardner, et al. (2005). "A comparison of RNA folding measures." BMC Bioinformatics 6: 241.

Freyhult, E., V. Moulton, et al. (2005). "Predicting RNA structure using mutual information." Appl Bioinformatics 4(1): 53-9.

Gegenheimer, P. and D. Apirion (1981). "Processing of procaryotic ribonucleic acid." Microbiol Rev 45(4): 502-41.

Gralla, J. and C. DeLisi (1974). "mRNA is expected to form stable secondary structures." Nature 248(446): 330-2.

Graveley, B. R., K. J. Hertel, et al. (1998). "A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers." EMBO J 17(22): 6747-56.

Grossberg, S., D. Levine, et al. (1987). "Predictive regulation of associative learning in a neural network by reinforcement and attentive feedback." Int J Neurol 21-22: 83-104.

Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." Cell 130(1): 77-88.

Gultyaev, A. P., F. H. van Batenburg, et al. (1995). "The computer simulation of RNA folding pathways using a genetic algorithm." J Mol Biol 250(1): 37-51.

Hermann, T. and D. J. Patel (2000). "RNA bulges as architectural and recognition motifs." Structure 8(3): R47-54.

Hewish, D. R. (1976). "DNA replication in eukaryotes: a model for the specific involvement of chromatin subunits." Nucleic Acids Res 3(1): 69-78.

Hiller, M., Z. Zhang, et al. (2007). "Pre-mRNA secondary structures influence exon recognition." PLoS Genet 3(11): e204.

Hokin, L. E. and M. R. Hokin (1954). "Ribonuclease and ribonucleic acid synthesis." J Histochem Cytochem 2(5): 395-400.

Holbrook, S. R. (2005). "RNA structure: the long and the short of it." Curr Opin Struct Biol 15(3): 302-8.

Holley, L. And M. Karplus (1989) Protein Secondary Structure Prediction with a Neural Network. Proc. Natl. Acad. Sci. USA Vol. 86 pp. 152-156.

Hou, Y. and S. Lin (2009). "Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes." PLoS One 4(9): e6978.

Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." Mol Biol Evol 2(1): 13-34.

Jacobson, A. B., R. Arora, et al. (1998). "Structural plasticity in RNA and its role in the regulation of protein translation in coliphage Q beta." J Mol Biol 275(4): 589-600.

Jia, M., L. Luo, et al. (2004). "Statistical correlation between protein secondary structure and messenger RNA stem-loop structure." Biopolymers 73(1): 16-26.

Katz, L. and C. B. Burge (2003). "Widespread selection for local RNA secondary structure in coding regions of bacterial genes." Genome Res 13(9): 2042-51.

Kornberg, A. (1988). "DNA replication." Biochim Biophys Acta 951(2-3): 235-9.

Kraemer, E., Wang, J, Guo, J., Hopkins, S., Arnold, J. (2001). An analysis of gene-finding programs for Neurospora crassa, Bioinformatics 17:901-12.

Kunkel, T. A. (1992). "DNA replication fidelity." J Biol Chem 267(26): 18251-4.

Kushner, S. R. (2002). "mRNA decay in Escherichia coli comes of age." J Bacteriol 184(17): 4658-65; discussion 4657.

Lagunez-Otero, J. and E. N. Trifonov (1992). "mRNA periodical infrastructure complementary to the proof-reading site in the ribosome." J Biomol Struct Dyn 10(3): 455-64.

Langridge, R., W. E. Seeds, et al. (1957). "Molecular structure of deoxyribonucleic acid (DNA)." J Biophys Biochem Cytol 3(5): 767-78.

Le, S. Y. and J. V. Maizel, Jr. (1989). "A method for assessing the statistical significance of RNA folding." J Theor Biol 138(4): 495-510.

Lesk, A.M. (1974) "A Combinatorial study of the effects of admitting non-Watson-Crick base pairing and of base composition on the helix forming potential of polynucleotides of random sequences." J Theor Biol 44. 7-17.

Li, J., S. Iwamoto, et al. (1997). "Dinucleotide repeat in the 3' flanking region provides a clue to the molecular evolution of the Duffy gene." Hum Genet 99(5): 573-7.

Liu, H. X., G. J. Goodall, et al. (1995). "Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in Nicotiana plumbaginifolia." EMBO J 14(2): 377-88.

Loeb, L. A., C. F. Springgate, et al. (1974). "Errors in DNA replication as a basis of malignant changes." Cancer Res 34(9): 2311-21.

Luo, L., M. Jia, et al. (2004). "Protein structure preference, tRNA copy number, and mRNA stem/loop content." Biopolymers 74(6): 432-47.

Mahen, E. M., P. Y. Watson, et al. (2010). "mRNA secondary structures fold sequentially but exchange rapidly in vivo." PLoS Biol 8(2): e1000307.

Mathews, D. H., A. R. Banerjee, et al. (1997). "Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable

element." RNA 3(1): 1-16.

Mathews, D. H., M. D. Disney, et al. (2004). "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." Proc Natl Acad Sci U S A 101(19): 7287-92.

Mathews, D. H., J. Sabina, et al. (1999). "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." J Mol Biol 288(5): 911-40.

McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." Biopolymers 29(6-7): 1105-19.

Meyer, I. M. and I. Miklos (2005). "Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs." Nucleic Acids Res 33(19): 6338-48.

Mita, K., S. Ichimura, et al. (1988). "Specific codon usage pattern and its implications on the secondary structure of silk fibroin mRNA." J Mol Biol 203(4): 917-25.

Montange, R. K. and R. T. Batey (2008). "Riboswitches: emerging themes in RNA structure and function." Annu Rev Biophys 37: 117-33.

Namy, O., S. J. Moran, et al. (2006). "A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting." Nature 441(7090): 244-7.

Nowakowski, J., P. J. Shim, et al. (1999). "Crystal structure of an 82-nucleotide RNA-DNA complex formed by the 10-23 DNA enzyme." Nat Struct Biol 6(2): 151-6.

Ogura, H. Agata, H., Xie, M., Odaka, T., and H. Furutani (1997) A Study of Learning Splice Sites of DNA Sequence by Neural Networks. Comput. Biol. Med. Vol. 27 No. 1pp. 67-75.

Organ, C. L., A. M. Shedlock, et al. (2007). "Origin of avian genome size and structure in non-avian dinosaurs." Nature 446(7132): 180-4.

Pace, N. R. (1973). "Structure and synthesis of the ribosomal ribonucleic acid of

prokaryotes." Bacteriol Rev 37(4): 562-603.

Pipas, J. M. and J. E. McMahon (1975). "Method for predicting RNA secondary structure." Proc Natl Acad Sci U S A 72(6): 2017-21.

Rho, J. H. and J. Bonner (1961). "The site of ribonucleic acid synthesis in the isolated nucleus." Proc Natl Acad Sci U S A 47: 1611-9.

Rich, A. and J. D. Watson (1954). "Some Relations between DNA and Rna." Proc Natl Acad Sci U S A 40(8): 759-64.

Ron, E. Z., R. E. Kohler, et al. (1966). "Increased stability of polysomes in an Escherichia coli mutant with relaxed control of RNA synthesis." Proc Natl Acad Sci U S A 56(2): 471-5.

Rosset, R. and R. Monier (1963). "[Apropos of the presence of weak molecular weight RNA in the ribosomes of Escherichia Coli.]." Biochim Biophys Acta 68: 653-6.

Sankaranarayanan, R., A. C. Dock-Bregeon, et al. (1999). "The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site." Cell 97(3): 371-81.

Schneider, A. and D. Ebert (2004). "Covariation of mitochondrial genome size with gene lengths: evidence for gene length reduction during mitochondrial evolution." J Mol Evol 59(1): 90-6.

Schroeder, S., J. Kim, et al. (1996). "G.A and U.U mismatches can stabilize RNA internal loops of three nucleotides." Biochemistry 35(50): 16105-9.

Seffens, W. and D. Digby (1999). "mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences." Nucleic Acids Res 27(7): 1578-84.

Serra, M. J., T. W. Barnes, et al. (1997). "Improved parameters for the prediction of RNA hairpin stability." Biochemistry 36(16): 4844-51.

Shabalina, S. A., A. Y. Ogurtsov, et al. (2006). "A periodic pattern of mRNA

secondary structure created by the genetic code." Nucleic Acids Res 34(8): 2428-37.

Shigeura, H. T. and E. Chargaff (1958). "Studies of the dynamics of ribonucleic acid formation." J Biol Chem 233(1): 197-202.

Skrisovska, L., C. F. Bourgeois, et al. (2007). "The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction." EMBO Rep 8(4): 372-9.

Smith, D. J., C. C. Query, et al. (2008). ""Nought may endure but mutability": spliceosome dynamics and the regulation of splicing." Mol Cell 30(6): 657-66.

Snyder, E. E. and G. D. Stormo 1993 "Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks" Nuc. Acids Res. 21:607-613.

Sussman, J. L., S. R. Holbrook, et al. (1978). "Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement." J Mol Biol 123(4): 607-30.

Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nat Rev Genet 9(6): 465-76.

Szymanski, M., M. Z. Barciszewska, et al. (2000). "5S ribosomal RNA database Y2K." Nucleic Acids Res 28(1): 166-7.

Taylor, M. M., J. E. Glasgow, et al. (1967). "Sedimentation coefficients of RNA from 70S and 80S ribosomes." Proc Natl Acad Sci U S A 57(1): 164-9.

Taylor, M. M. and R. Storck (1964). "Uniqueness of Bacterial Ribosomes." Proc Natl Acad Sci U S A 52: 958-65.

Tinoco, I., Jr., O. C. Uhlenbeck, et al. (1971). "Estimation of secondary structure in ribonucleic acids." Nature 230(5293): 362-7.

Tipper, D. J. and K. A. Bostian (1984). "Double-stranded ribonucleic acid killer systems in yeasts." Microbiol Rev 48(2): 125-56.

van Batenburg, F. H., A. P. Gultyaev, et al. (1995). "An APL-programmed genetic algorithm for the prediction of RNA secondary structure." J Theor Biol 174(3): 269-80.

von Heijne, G., L. Nilsson, et al. (1978). "Models for mRNA translation: theory versus experiment." Eur J Biochem 92(2): 397-402.

Wang, L. and S. R. Wessler (2001). "Role of mRNA secondary structure in translational repression of the maize transcriptional activator Lc(1,2)." Plant Physiol 125(3): 1380-7.

Waterman, M.S., (1978). "Secondary structure of single-stranded nucleic acids." Adv. Math Suppl. Study., 1. 167-212

Watson, J. D. and F. H. Crick (1953). "The structure of DNA." Cold Spring Harb Symp Quant Biol 18: 123-31.

White, H. B., 3rd, B. E. Laux, et al. (1972). "Messenger RNA structure: compatibility of hairpin loops with protein sequence." Science 175(27): 1264-6.

Woodson, S. A. (2008). "RNA folding and ribosome assembly." Curr Opin Chem Biol 12(6): 667-73.

Workman, C. and A. Krogh (1999). "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution." Nucleic Acids Res 27(24): 4816-22.

Wu, C. H. (1997). "Artificial neural networks for molecular sequence analysis." Comput Chem 21(4): 237-56.

Zama, M. (1999). "Correlation between mRNA structure of the coding region and translational pauses." Nucleic Acids Symp Ser(42): 81-2.

Zhang, J. (2000). "Protein-length distributions for the three domains of life." Trends Genet 16(3): 107-9.

Zuker, M. (1989). "Computer prediction of RNA structure." Methods Enzymol 180: 262-88.

Zuker, M. (1989). "On finding all suboptimal foldings of an RNA molecule." Science 244(4900): 48-52.

Zuker, M. (2000). "Calculating nucleic acid secondary structure." Curr Opin Struct Biol 10(3): 303-10.